4 Bayesian Theory

As we have seen, in frequentist statistics statements are made with reference to repetitions of the same experiment with parameters fixed. In Bayesian statistics, parameters are no longer regarded as fixed, but are themselves random variables. The probability distribution of the parameter values before taking data, the **prior distribution**, is updated to a probability distribution after taking data, the **posterior distribution**, through the likelihood of the observed data. This update is achieved through **Bayes' Theorem**. Bayesian inference attempts to say as much as possible about the unknown parameter distribution based on the observed data only, without reference to future repetitions of the same experiment. Bayesian posteriors are probability distributions on the unknown parameter and can be interpreted and manipulated in that way, as statements about the relative probability that the parameter takes different values.

The derivation of Bayes' theorem is a mathematical result that follows from the definition of conditional probability, as we will see below, but it is how this result is applied to interpret data, and the philosophical distinction in the interpretation of the parameter values that distinguishes the frequentist and Bayesian approach. Typically, in any given observation, the actual parameter values that led to the generation of the observed data are fixed, not random, but the Bayesian interpretation is that you can never by sure of what the unknown parameter is, and so it is appropriate to consider it to be a random variable. In many cases you will not be able to repeat a particular experiment. Gravitational wave observations are a good example of this — we cannot choose what events occur in the Universe, so every observed event is a unique, non-repeatable, experiment. In such contexts, the frequentist approach of referencing theoretical repetitions cannot really be seen as representative of reality. In cases where it is possible to repeat an experiment with the unknown parameters fixed, the Bayesian posterior converges to the true parameter value asymptotically and so can still be used to represent the current level of uncertainty in the parameter.

Frequentist concepts such as significance and hypothesis testing have been incorporated into the Bayesian framework, but the interpretation in the latter context is not always clean. It is therefore useful to have familiarity with both sets of tools to be fully quipped to handle any kind of data analysis problem.

4.1 Conditional probability

It is often the case that a process generates more than one potentially measurable random output, but only a subset of these are measurable. If the variables are independent then measuring one would not provide any information about the others, but when there are inter-dependencies the observation of a random variable can provide information about other variables with which it is correlated. For example, suppose we have a bag containing 100 balsa, of which 10 are red and stripy, 20 are blue and stripy, 30 are red and spotted and 40 are blue and spotted. In total there are 30 stripy balls out of the 100 and therefore the probability that a randomly chosen ball is stripy is 3/10. However, out of the 40 red balls there are only 10 that are stripy, and so if we have observed that the ball is red the probability that it is also stripy is now 1/4.

The **conditional probability** of an event A, given some other event B is defined as

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

In other words, this is the fraction that both A and B occur, our of all the times that B occurs. This can be rewritten in two different ways by interchanging A and B

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A).$$

Rearranging this identity we obtain **Bayes' Theorem**

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

4.2 Bayesian inference

Bayes' Theorem is a mathematical identity, but it becomes philosophically distinct from frequentist approaches when it is applied to inference. In Bayesian inference, the event A is taken to be an observation of data, $\mathbf{x}$, and the event B is taken to be the value of some unknown parameters, $\vec{\theta}$, characterising the system being observed. Bayes' Theorem becomes

$$p(\vec{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta})}{p(\mathbf{x})}.$$

In this context $p(\mathbf{x}|\vec{\theta})$ is the likelihood (the same function of data and parameters as in the frequentist case), $p(\vec{\theta})$ is the **prior** distribution of source parameter values, $p(\vec{\theta}|\mathbf{x})$ is the **posterior** distribution on the source parameter values and $p(\mathbf{x})$ is the **evidence** for the model under consideration. In a parameter estimation context, the evidence, which does not depend on parameter values, is a normalisation constant that can be ignored. However, it plays an important role in Bayesian hypothesis testing, which will be discussed in section 4.6.

Example: Medical testing We suppose that a medical test for a disease is 95% effective but has a 1% false alarm rate and the prevalence of the disease in the population is 0.5%. You test positive for the disease. What is the probability you do in fact have it?

The term "95% effective" means that if you have the disease the test gives a positive result 95% of the time. The term 1% false alarm rate means that if you do not have the disease you test positive 1% of the time. We can now apply Bayes theorem with data $\mathbf{x}$ = 'positive test' and parameter θ ='disease status' taking values 'infected' or 'not infected'. The likelihood is

$$p(\text{positive}|\text{infected}) = 0.95, \qquad p(\text{positive}|\text{not infected}) = 0.01.$$

The prior is based on the known prevalence in the population

$$p(\text{infected}) = 1 - p(\text{not infected}) = 0.005.$$

The posterior is then

$$\begin{aligned} p(\text{infected}|\text{positive}) &= \frac{p(\text{positive}|\text{infected})p(\text{infected})}{p(\text{positive}|\text{infected})p(\text{infected}) + p(\text{positive}|\text{not infected})p(\text{not infected})} \\ &= \frac{0.95 * 0.005}{0.95 * 0.005 + 0.01 * 0.995} = 0.323. \end{aligned} \tag{59}$$

So you are more likely not to be infected than to be infected if you get a positive test result. The solution is to get a second opinion. If you take a second (independent) test and it is also positive your posterior probably of being infected is now

$$p(\text{infected}|\text{2nd positive}) = \frac{0.95 * 0.323}{0.95 * 0.323 + 0.01 * 0.677} = 0.978 = \frac{0.95^2 * 0.005}{0.95^2 * 0.005 + 0.01^2 * 0.995}.$$

The first of these two results follows from using the posterior from the first test as a prior for the second. The second result follows from regarding the observed data as "two independent positive tests".

Example: Blood evidence Based on other evidence, a detective is 50% sure that a particular suspect has committed a murder. Then new evidence comes to light. A small amount of blood, of type B, is found at the scene. This is not the victim's blood type, but it is the blood type of the suspect. Such a blood type has a prevalence of 2% in the population. What is the detective's confidence in the guilt of the suspect in light of this new evidence?

The likelihood is

$$p(\text{type B blood}|\text{guilty}) = 1, \qquad p(\text{type B blood}|\text{not guilty}) = 0.02.$$

The prior is $p(\text{guilty}) = 0.5$ and so the posterior is

$$\begin{aligned} p(\text{guilty}|\text{type B blood}) &= \frac{p(\text{type B blood}|\text{guilty})p(\text{guilty})}{p(\text{type B blood}|\text{guilty})p(\text{guilty}) + p(\text{type B blood}|\text{not guilty})p(\text{not guilty})} \\ &= \frac{0.5}{0.5 + 0.01} = 0.98. \end{aligned} \tag{60}$$

4.3 Choice of prior

The prior plays a key role in Bayesian parameter inference. It expresses the current state of our understanding about parameter values, and it is updated to the posterior using data via the likelihood. Mathematically, the prior represents the distribution of the unknown parameter value in nature, but usually this is not known. In that case, the prior reflects the current state of knowledge about the parameter values, which may come from previous experiments or expert opinion or not be known.

4.3.1 Informative/expert priors

If information is available, it is appropriate to use informative priors. For example, if previous measurements have been made of a quantity it is reasonable to use the posterior from those measurements as a prior for the next measurement, as we saw in the medical test example above. Alternatively, even if a measurement has not been made directly, "experts" may be able to give a reasonable range or distribution for the parameter based on experience in other situations. One criticism that is often levelled at Bayesian inference is that the result can depend on the assumed prior. However, the Bayesian response is that this is desired behaviour — if we have additional information from prior knowledge, then it is the correct thing to do to include that in our conclusions based on subsequent observed data.

The process of constructing a prior based on the opinion of experts is known as **elicitation**. Sometimes, elicitation may result in different priors from different experts. In that case a **mixture prior** can be constructed

$$p(\vec{\theta}) = \sum_{j=1}^{J} \omega_j p_j(\vec{\theta})$$

where j labels which of the J experts we are referring to, $p_j(\vec{\theta})$ is the prior elicited from that expert, and ω_j is the weight given to that expert (or set of experts).

If the prior is based on the posterior from previous observations it is normally clear how to fold this in. If the prior comes from expert opinion, it may be possible to use this in several different ways. In that case, care must be taken to be as conservative as is reasonably possible in the use of that prior information, to avoid making conclusions form the data that are too strong.

4.3.2 Conjugate priors

It is convenient to choose a form for the prior that ensures the posterior takes the same form. In such situations, the posterior from an experiment can be directly be used as a prior for the next experiment and so on. Such a prior is called **conjugate**.

Definition: A family of distributions, $\mathcal{F}$, is **conjugate** to a family of sampling distributions, $\mathcal{P}$, if, whenever the prior belongs to the family $\mathcal{F}$, the posterior belongs to the same family, for any number and value of observations from $\mathcal{P}$.

The form of the conjugate prior depends on the nature of the probability distribution, $\mathcal{P}$, from which the observed data is drawn. This gives rise to a number of conjugate families. In particular, any distribution in the exponential family

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \;\forall x, \vec{\theta}$$

has a conjugate prior in the exponential family of the form

$$p(\vec{\theta}|\vec{\chi},\nu) = p(\vec{\chi},\nu)\exp\left[\vec{\theta}^T\vec{\chi} - \nu A(\vec{\theta})\right] \tag{61}$$

where ν and $\vec{\chi}$ are the hyperparameters of the prior distribution.

A full list of conjugate priors can be found in the conjugate prior entry on wikipedia, but the three most widely used are the Beta-Binomial, Poisson-Gamma and Normal-Normal families, and we will discuss these further here.

Beta-Binomial model Suppose our observed data $\mathbf{X}$ ~Bin(n,p) with likelihood

$$p(\mathbf{x}|p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}.$$

The conjugate prior is the Beta(a,b) distribution with density

$$p(p) = \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}.$$

Observing binomial distributed data and using the Beta prior gives a posterior

$$\begin{aligned} p(p \mid x) &\propto p(x \mid p)p(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \\ &\propto p^{a+x-1}(1-p)^{b+n-x-1}. \end{aligned}$$

So the posterior is also a Beta distribution

$$p(p \mid x) = \text{Beta}(a+x, b+n-x).$$

The mean and variance of a Beta(a, b) distribution are

$$\mathbb{E}(\mathbf{X}) = \frac{a}{a+b}, \qquad \text{var}(\mathbf{X}) = \frac{ab}{(a+b)^2(a+b+1)}.$$

The posterior mean is therefore

$$\mathbb{E}(p|x) = \frac{a+x}{a+b+n}$$

which we compare to the mean in the observed data of x/n. One interpretation of the prior data is that it represents having observed $a-1$ events in $a+b-2$ previous trials. If a and b are kept fixed and $n, x \to \infty$ the posterior mean tends to the maximum likelihood estimator x/n and the posterior variance tends to zero.

Poisson-Gamma model Suppose now that we are observing data, $X_1, \ldots, X_n$, from a Poisson distribution, $\mathbf{X} \sim$Pois(λ), with likelihood

$$p(\mathbf{x} \mid \lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\}.$$

The conjugate prior is the Gamma(m, μ) distribution

$$p(\lambda | m, \mu) = \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda},$$

which has mean m/μ and variance m/μ^2. With this prior the posterior is

$$\begin{aligned} p(\lambda \mid \mathbf{x}) &\propto p(\mathbf{x}|\lambda)p(\lambda) \\ &= \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\} \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda} \\ &\propto \mathrm{e}^{-n\lambda - \mu\lambda} \lambda^{\sum_{i=1}^{n} x_i + m - 1} \\ &\propto \text{Gamma}(m + n\bar{x}, \mu + n). \end{aligned} \tag{62}$$

The posterior mean can be seen to equal

$$\mathbb{E}(p(\lambda \mid \mathbf{x})) = \frac{m+n\bar{x}}{m+n} = \bar{x}\left(\frac{n}{n+m}\right) + \frac{m}{\mu}\left(1 - \frac{n}{n+m}\right),$$

i.e., it is a compromise between the prior mean, m/μ, and the maximum likelihood estimator $\bar{x}$. As the number of samples increases, more weight is placed on the data and less on the prior, as expected.

Normal-Normal/Normal-Gamma model Now we consider $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, and likelihood

$$p(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

We assume first that σ^2 is known. The conjugate prior in this case is the Normal distribution, $N(\mu_0, \sigma_0^2)$,

$$p(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right].$$

The posterior is

$$\begin{aligned} p(\mu \mid \mathbf{x}, \sigma^2) &\propto p(\mathbf{x} \mid \mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\sigma_0^2}\left[\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)\right]\right\}, \end{aligned}$$

which can be recognized as a $N(\mu_n, \sigma_n^2)$ distribution, where

$$\mu_n = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}. \tag{63}$$

Writing these results in terms of $\tau = 1/\sigma^2$, which is called the **precision** of the Normal distribution we can see

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{y}$$

so once again the posterior mean is a balance between the prior mean and the sample mean, with the relative weighting determined by both the number of observations and the relative precision of the observations and the prior.

If we suppose that μ is known (which is an unrealistic assumption in practice), but the variance is uncertain, then we can obtain a conjugate prior by using a Gamma(a, b) prior on the precision

$$p(\tau|a, b) \propto \tau^{a-1}e^{-b\tau}$$

and obtain the posterior

$$\begin{aligned} p(\tau \mid \mathbf{x}, \mu) &\propto p(\mathbf{x} \mid \mu, \tau)p(\tau|a, b) \\ &\propto \tau^{n/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}\tau^{a-1}e^{-b\tau} \\ &= \tau^{a+n/2-1}\exp\left\{-\tau\left(b + \frac{1}{2}\sum_i (x_i - \mu)^2\right)\right\} \\ &\sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \end{aligned}$$

It is common practice to take the limit in which a and b are both very small and then the posterior becomes

$$p(\tau \mid \mathbf{x}, \mu) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \quad \Rightarrow \quad \mathbb{E}\left[\tau \mid \mathbf{x}, \mu\right] = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-1},$$

so the posterior expectation of the precision is approximately the same as the (frequentist) sample precision (up to a factor of $n/(n-1)$).

Finally we assume that both μ and σ^2 are unknown. It would be reasonable to just multiply together the two previous priors, but this does not result in a conjugate prior, essentially because the posterior on μ in the first case depends on the known variance σ^2. However, we can find a correlated conjugate prior (writing $\tau = 1/\sigma^2$ as before) by writing

$$\mu \sim N(\mu_0, 1/(n_0\tau)), \quad \tau \sim \text{Gamma}(a, b),$$

or, explicitly,

$$p(\mu, \tau | \mu_0, n_0, a, b) \propto \left(\frac{n_0\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{n_0\tau}{2}(\mu-\mu_0)^2\right]\tau^{a-1}e^{-b\tau}.$$

The posterior on μ, conditioned on τ, $p(\mu|\tau, \mathbf{x})$, is given by the same expression as before

$$p(\mu|\tau, \mathbf{x}) \sim N\left(\frac{n_0\mu_0 + n\bar{x}}{n_0+n}, \frac{1}{(n_0+n)\tau}\right).$$

The posterior on τ can be found by considering the combined posterior, being careful not to drop any terms that depend on μ or τ

$$\begin{aligned} p(\mu, \tau|\mathbf{x}) &\propto \sqrt{\tau}\exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\tau^{\frac{n}{2}}\exp\left[-\frac{n_0\tau}{2}(\mu-\mu_0)^2\right]\tau^{a-1}e^{-b\tau} \\ &= \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b - \frac{(n\bar{x}+n_0\mu_0)^2}{2(n+n_0)} + \frac{1}{2}n_0\mu_0^2 + \frac{1}{2}\sum x_i^2\right)\tau\right]\times \\ &\times\left(\sqrt{\frac{(n+n_0)\tau}{2\pi}}\exp\left[-\frac{(n+n_0)\tau}{2}\left(\mu - \frac{(n\bar{x}+n_0\mu_0)}{n+n_0}\right)^2\right]\right). \end{aligned} \tag{64}$$

If we now marginalise over μ, the round bracketed term on the final line integrates to a constant, independent of τ, and the term inside the exponent on the penultimate line can be simplified to obtain

$$\begin{aligned} p(\tau|\mathbf{x}) &\propto \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\mu_0-\bar{x})^2\right)\tau\right] \\ \Rightarrow p(\tau|\mathbf{x}) &\sim \text{Gamma}\left(a+\frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\mu_0-\bar{x})^2\right). \end{aligned} \tag{65}$$

And so this is also a conjugate prior model, called the Normal-Gamma model.

4.3.3 Using expert information with conjugate priors

If expert prior information is in the form of a posterior from a previous experiment the form of the distribution is fixed. However, in other circumstances it can be possible to express the prior information in the form of a particular choice of parameters for a conjugate prior. This is most clearly seen with an example.

Example: Consider a drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates, p, between 0.2 and 0.6 could be feasible. We plan to observe the response rate in n patients and want to infer a posterior on p. Propose a suitable conjugate prior for p based on the available information.

A response rate between 0.2 and 0.6 could be used to set a uniform prior in that range. However, this is not conjugate to the binomial distribution that determines the observed data. Therefore, it would be better to use a conjugate prior. A $U[0.2, 0.6]$ distribution has mean 0.4 and standard deviation of 0.1. We can find a Beta distribution that has the same mean and standard deviation. Rearranging the equations given earlier we deduce $\text{Beta}(a = 9.2, b = 13.8)$ has the desired mean and variance. This prior is conjugate and reflects the expert opinion as regards the expected response rate for the drug. Suppose now we observe $n = 20$ patients and $x = 15$ respond positively. The posterior is then $\text{Beta}(9.2 + 15, 13.8 + 5) = \text{Beta}(24.2, 18.8)$. The prior, (scaled) likelihood and posterior are illustrated in Figure 4.

4.3.4 Mixture priors

The use of a conjugate prior can be somewhat restrictive as there is limited flexibility within the prior family. However, one way to get around this is by using **mixture priors**. A mixture prior is of the form

$$p(\vec{\theta}) = \sum_{j=i}^{J} \pi_j p(\vec{\theta} \mid \vec{\psi}_j), \quad \sum_{j=1}^{J} \pi_j = 1. \tag{66}$$

Here $\{\pi_j\}$ are called the mixture weights and it is assumed that the hyperparameters, ψ_j, are different in each component. If the mixture components are all drawn from the conjugate prior family, then the mixture prior is also conjugate.

Example: Beta-Binomial mixture prior Suppose $X \sim \text{Bin}(n, p)$ and we use a prior on p that is a mixture distribution

$$p(p|a_1, b_1, a_2, b_2) = \pi\text{Beta}(a_1, b_1) + (1 - \pi)\text{Beta}(a_2, b_2).$$

What is the posterior distribution for p?

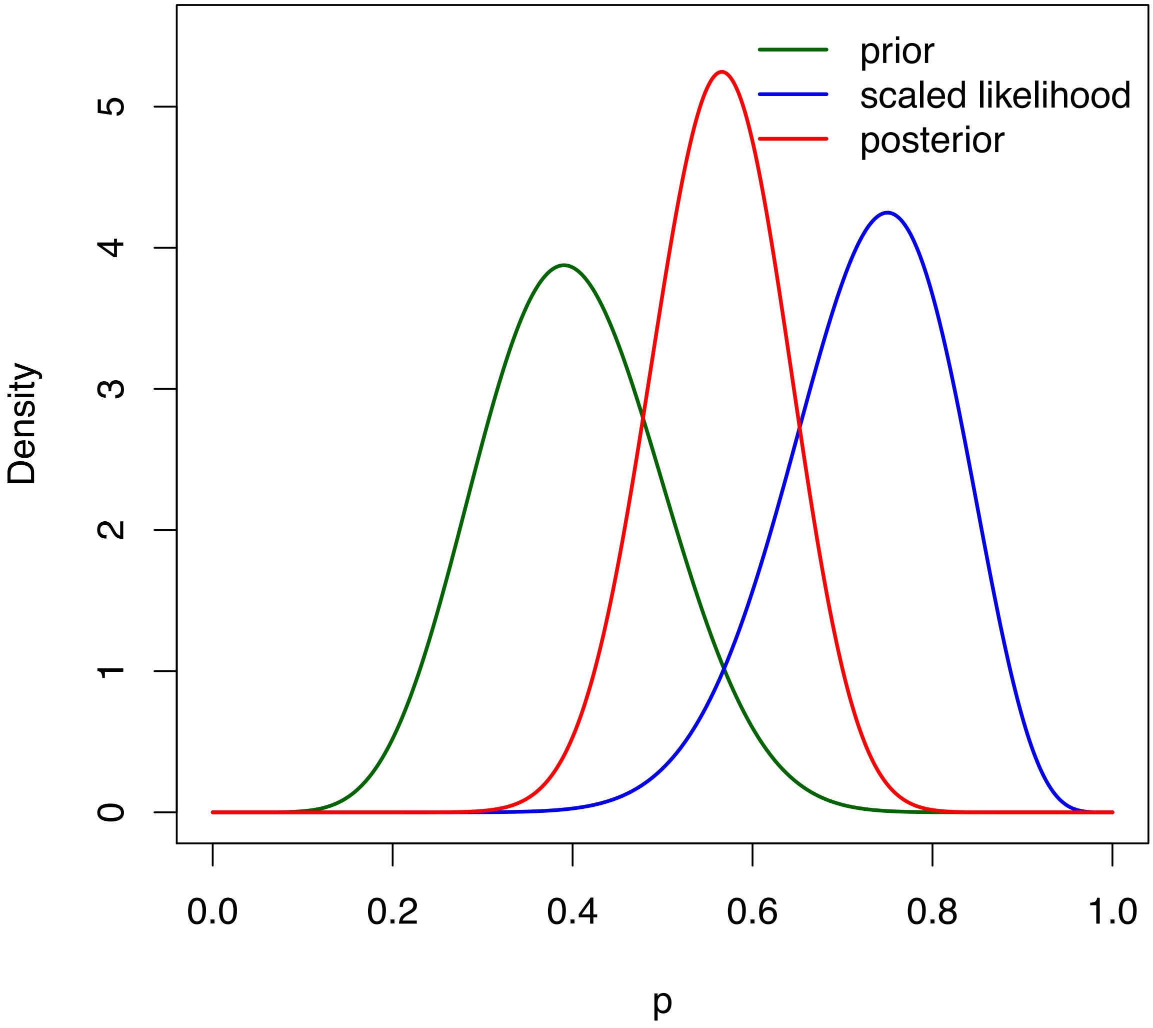

Figure 4: Conjugate prior, Beta$(9.2, 13.8)$, likelihood, Bin$(20, p)$, and posterior, Beta$(24.2, 18.8)$ for the drug response problem described in the text. The likelihood has been rescaled to ensure it has a similar height to the prior and posterior distributions.

Solution: We find the posterior as follows

$$
\begin{aligned}
p(p \mid x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \left\{ \pi \frac{1}{B(a_1,b_1)} p^{a_1-1}(1-p)^{b_1-1} + (1-\pi)\frac{1}{B(a_2,b_2)} p^{a_2-1}(1-p)^{b_2-1} \right\} \\
&\propto \pi \frac{1}{B(a_1,b_1)} p^{a_1+x-1}(1-p)^{b_1+n-x-1} + (1-\pi)\frac{1}{B(a_2,b_2)} p^{a_2+x-1}(1-p)^{b_2+n-x-1} \\
&= \pi \frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)} \frac{1}{B(a_1+x,b_1+n-x)} p^{a_1+x-1}(1-p)^{b_1+n-x-1} \\
&\quad + (1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)} \frac{1}{B(a_2+x,b_2+n-x)} p^{a_2+x-1}(1-p)^{b_2+n-x-1} \\
&= \pi \frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)} \mathrm{Beta}(p \mid a_1+x, b_1+n-x) \\
&\quad + (1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)} \mathrm{Beta}(p \mid a_2+x, b_2+n-x).
\end{aligned}
$$

We finish by normalising the weights to obtain

$$
p \mid x \sim \omega_1 \mathrm{Beta}(p \mid a_1+x, b_1+n-x) + (1-\omega_1)\mathrm{Beta}(p \mid a_2+x, b_2+n-x)
$$

with

$$
\omega_1 = \pi \frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)} \left(\pi \frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)} + (1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)} \right)^{-1}
$$

So the posterior is also a mixture of Beta distributions.

4.3.5 Jeffreys prior

If we do not have any prior information, it is normal to use an "uninformative" prior, i.e., a prior that assumes as little as possible about the parameter values. It is common to use uniform priors as uninformative priors, so that the posterior basically corresponds to the likelihood of the data. This is approach taken for many parameters in parameter estimation of gravitational wave data and was in fact the approach that Bayes himself advocated. However, uniform priors are not invariant under re-parameterisation. If one is ignorant about the value of θ, one is also ignorant about the value of θ^2 or any other function of θ. Therefore, any uninformative prior should induce the same form of uninformative prior on any other variables defined by transformation. Jeffreys (1961) proposed a class of priors that are invariant under re-parameterisations. By identifying the probability density with a metric on parameter space he argued that the prior should take the form $[\det(g_{ij})]^{1/2}$ where the metric

$$
g_{ij}(\vec{\theta}) = \frac{1}{f(\vec{\theta})} \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j}.
$$

This would lead to an invariant prior for any scalar function $f(\vec{\theta})$. Jeffreys advocated the use of the likelihood, which introduces a data dependence into the expression, that can be eliminated by taking the expectation over realisations of the data. This procedure leads to **Jeffreys prior** which is

$$
p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right]
$$

for $l = \log p(\mathbf{x}|\vec{\theta})$ the log-likelihood is the Fisher information matrix.

Jeffreys prior is “uninformative” because it can be interpreted as being as close as possible to the likelihood function and it is invariant under re-parameterisation. However, it is rarely a member of the conjugate family of distributions or of some other convenient form which is why it is not always convenient to use it in practice. Note also that the Jeffreys prior is not always **proper**, i.e., it does not always have a finite integral and therefore may not be normalisable.

Example: Poisson distribution For a single observation, x, from the Poisson(λ) distribution with pmf

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

we have

$$\frac{\partial \log p}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \frac{\partial^2 \log p}{\partial \lambda^2} = -\frac{x}{\lambda^2} \quad \Rightarrow \quad I(\lambda) \equiv \mathbb{E}\left[-\frac{\partial^2 \log p}{\partial \lambda^2}\right] = \frac{1}{\lambda}.$$

The Jeffreys prior for the Poisson distribution is therefore $p(\lambda) \propto 1/\sqrt{\lambda}$. This is an example of an **improper** prior, since it cannot be normalised to integrate to 1 unless the range of rates is restricted.

4.4 Posterior summary statistics

The result of a Bayesian inference calculation is a probability distribution, the full posterior probability distribution of the parameters, $p(\vec{\theta}|\mathbf{x})$. This is not only difficult to calculate in many cases, it is also unwieldy to manipulate and so it is common to use quantities that summarise the properties of the distribution. These are all of the summary statistics that we encountered in the first chapter of the course.

4.4.1 Point estimates

To obtain point estimates of a parameter value, θ_1 say, one typically works with the **marginalised** distribution for that parameter, defined by

$$p_{\mathrm{marg}}(\theta_1|\mathbf{x}) = \int p(\vec{\theta}|\mathbf{x}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_m.$$

From this marginal distribution, we can evaluate the **posterior mean**

$$\mu = \int_{-\infty}^{\infty} \theta_1 p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior median**, m, defined such that

$$\int_{-\infty}^{m} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1 = 0.5 = \int_{m}^{\infty} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior mode**

$$M = \mathrm{argmax}\ p_{\mathrm{marg}}(\theta_1|\mathbf{x}).$$

The posterior mean and mode can be defined unambiguously over the full distribution as well. The posterior mean is the same whether computed over the marginal distribution or the full distribution, but the mode typically changes. The median is not unambiguously defined on the whole distribution, as there are infinitely many ways to partition the full parameter space into equal probability subsets.

4.4.2 Credible intervals

To move beyond point estimates, it is natural to want to describe ranges in which parameter values are estimated to lie. The Bayesian equivalent of a frequentist confidence interval is a **credible interval**. This is defined as

Definition: An interval (a, b) is a $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_a^b p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

A **credible region** can be defined in a similar way. This is any partition of parameter space that contains $100(1-\alpha)\%$ of the total posterior probability. Clearly credible intervals and regions are not unique, but there are two types of credible interval that are commonly used.

Definition: An interval (a, b) is a **symmetric** $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_{-\infty}^a p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = \frac{\alpha}{2} = \int_b^\infty p_{\rm marg}(\theta_1|\mathbf{x})\mathrm{d}\theta_1.$$

Definition: An interval (a, b) is a $100(1-\alpha)\%$ **highest posterior density (HPD) interval** for θ_1 if

1. $[a, b]$ is a $100(1-\alpha)\%$ credible interval for θ_1;
2. for all $\theta \in [a, b]$ and $\theta' \notin [a, b]$ we have $p_{\rm marg}(\theta|\mathbf{x}) \geq p_{\rm marg}(\theta'|\mathbf{x})$.

Credible intervals are more intuitive than confidence intervals as they make an explicit statement about the probability that the parameter takes values in the range, rather than referencing an ensemble of similar experiments.

4.4.3 Posterior samples

Summary statistics provide a useful way to summarise and compare distributions, but they inevitably discard information. To retain full information about the parameters we need the full posterior. Often this cannot be written down in a simple analytic form, but it can be summarised by drawing a set of samples $\{\vec{\theta}_1, \dots, \vec{\theta}_M\}$ randomly from the posterior. Such samples can then be used to compute integrals over the posterior

$$\int f(\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta} \approx \frac{1}{M}\sum_{i=1}^M f(\vec{\theta}_i).$$

Most quantities that one might wish to compute from a posterior distribution can be expressed as integrals of this form, and so generation of such samples is the most complete way to represent posterior distributions. Efficient production of samples is non-trivial and will be the topic of the next chapter of these notes.

4.5 Interpreting summary statistics

4.5.1 Decision theory

The posterior mean, mode and median are all valid ways to summarise a posterior distribution. One way to motivate these (and other possible) choices is through **decision theory**. In decision theory, understanding which decision is the best is motivated by introducing a **loss function** which characterises the cost or penalty of making a particular decision. Formally we define various quantities

- The **sample space** $\mathcal{X}$ denotes the possible values for the observed data, $\mathbf{x}$.
- The **parameter space**, Ω_θ, denotes possible (unknown) states of nature (or parameter values characterising the true pdf of observed data sets).
- We define a **family of probability distributions**, $\{\mathbb{P}_\theta(x) : x \in \mathcal{X}, \theta \in \Omega_\theta\}$, which describe how the observed data is generated in the possible states of nature.
- The **action space**, $\mathcal{A}$, is the set of actions that an experimenter can take after observing data, e.g., reject or accept a null hypothesis, assign an estimate to the value of θ etc.
- The **loss function**, $L : \Omega_\theta \times \mathcal{A} \to \mathbb{R}$, is a mapping from the space of actions and parameters to the real numbers, such that $L(a, \theta)$ is the loss associated with taking the action a when the true state of nature is θ.
- The set of **decision rules**, $\mathcal{D}$, is a set of mappings from data to actions. Each element $d \in \mathcal{D}$ is a function $d : \mathcal{X} \to \mathcal{A}$ that associates a particular action with each possible observed data set.

For a parameter value $\theta \in \Omega_\theta$, the risk of a decision rule, d, is defined as

$$R(\theta, d) = \mathbb{E}_\theta L(\theta, d(X)) = \left\{ \begin{array}{ll} \sum_{x \in \mathcal{X}} L(\theta, d(x)) p(x; \theta) & \text{for discrete } X \\ \int_{\mathcal{X}} L(\theta, d(x)) p(x; \theta) \mathrm{d}x & \text{for continuous } \mathcal{X}. \end{array} \right.$$

In other words, the risk is the expected loss of a particular decision rule when the true value of the unknown parameter is θ. Note that this is fundamentally a frequentist concept, since the definition implicitly invokes the idea of repeated samples from the parameter space $\mathcal{X}$ and computes the average loss over these hypothetical repetitions. However, it is possible to extend these ideas to a Bayesian framework by defining a prior, $\pi(\theta)$, over the parameters of the distribution. The **Bayes risk** of a decision rule, d, is then defined as

$$r(\pi, d) = \int_{\theta \in \Omega_\theta} R(\theta, d) \pi(\theta) \mathrm{d}\theta,$$

or by a sum in the case of a discrete-valued probability distribution. A decision rule is **a Bayes rule** with respect to the prior $\pi(\cdot)$ if it minimizes the Bayes risk, i.e.,

$$r(\pi, d) = \inf_{d' \in \mathcal{D}} r(\pi, d') = m_\pi, \text{ say.}$$

Note that, as usual in a Bayesian context, the Bayes rule depends on the specification of the prior and therefore there will be infinitely many Bayes rules for any particular problem. A useful choice of prior is the one that is most conservative in its estimate of risk. This gives rise to the concept of **a least favourable prior**. The prior $\pi(\theta)$ is least favourable if, for any other prior $\pi'(\theta)$ we have

$$r(\pi, d_\pi) \geq r(\pi', d_{\pi'})$$

where d_π, $d_{\pi'}$ are the Bayes rules corresponding to $\pi(\cdot)$ and $\pi'(\cdot)$ respectively.

4.5.2 Bayes rules as minimizers of posterior expected loss

The Bayes risk can be written as

$$\begin{aligned} r(\pi, d) &= \int_{\Omega_\theta} R(\theta, d)\pi(\theta)\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(x|\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(\theta|x)p(x)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\mathcal{X}} p(x)\left\{\int_{\Omega_\theta} L(\theta, d(x))p(\theta|x)\mathrm{d}\theta\right\}\mathrm{d}x \end{aligned}$$

where the second line follows from the definition of the risk function and the third line follows by using Bayes' theorem to write $p(x|\theta)\pi(\theta) = p(\theta|x)p(x)$ in terms of the posterior $p(\theta|x)$ and the evidence $p(x)$. The Bayes rule minimizes the Bayes risk. We see that this minimum is achieved for a particular value of x by making the decision that minimizes the expression in curly brackets. This is the **expected posterior loss** associated with the observed x. This observation simplifies the calculation in many cases and also illustrates the general property of Bayesian procedures, namely that the decision depends only on the observed data and not on potential unobserved data sets.

We will illustrate this with four examples. In the first three examples, we are attempting to make a point estimate and so the decision is an assignment of the value of the parameter $d = \hat{\theta}$.

Example: Point estimation with squared error loss Suppose we want to make a point estimate of a parameter and we use a squared error loss function, $L(\theta, d) = (\theta - d)^2$. Find the Bayes rule.

Solution

The Bayes rule chooses $d(Y)$ to minimize

$$\int_{\Omega_\theta} (\theta - d)^2 p(\theta|y)\mathrm{d}\theta.$$

Differentiating with respect to d and setting this to zero gives

$$\int_{\Omega_\theta} (\theta - d)p(\theta|x)\mathrm{d}\theta = 0 \quad \Rightarrow \quad d = \int_{\Omega_\theta} \theta p(\theta|x)\mathrm{d}\theta.$$

In other words, the Bayes estimator of θ, with squared error loss, is the **posterior mean**.

Example: Point estimation with absolute magnitude error loss
Suppose we instead use the loss function $L(\theta, d) = |\theta - d|$. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$\int_{-\infty}^{d} (d-\theta)p(\theta|x)\mathrm{d}\theta + \int_{d}^{\infty} (\theta - d)p(\theta|x)\mathrm{d}\theta.$$

Setting the derivative with respect to d to zero now gives

$$\int_{-\infty}^{d} p(\theta|x)\mathrm{d}\theta - \int_{d}^{\infty} p(\theta|x)\mathrm{d}\theta = 0 \quad \Rightarrow \quad \int_{-\infty}^{d} p(\theta|x)\mathrm{d}\theta = \int_{d}^{\infty} p(\theta|x)\mathrm{d}\theta = \frac{1}{2}.$$

In other words, the Bayes estimator of θ, with absolute magnitude error loss, is the **posterior median**.

Example: Point estimation with delta-function gain
Suppose we instead use the loss function

$$L(\theta, d) = \left\{ \begin{array}{cl} -\delta(\theta - d) & \text{if } d = \theta \\ 0 & \text{if } d \neq \theta \end{array} \right. .$$

In other words, the loss is infinitely higher for any value except the correct one. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$-\int_{-\infty}^{\infty} \delta(\theta - d)p(\theta|x)\mathrm{d}\theta = -p(d|x).$$

The minimum loss is obtained by setting

$$d = \text{argmaxp(d|x)},$$

i.e., the posterior mode.

Example: Interval estimation
Suppose we have a loss function of the form

$$L(\theta, d) = \left\{ \begin{array}{ll} 0 & \text{if } |\theta - d| \leq \delta \\ 1 & \text{if } |\theta - d| > \delta \end{array} \right.$$

for specified $\delta > 0$. What is the Bayes rule?

Solution

The expected posterior loss in this case is the posterior probability that $|\theta - d| > \delta$. The interval that minimises this loss, among intervals of fixed length 2δ, is the interval that contains the highest posterior probability. This is called the *highest posterior density* interval.

We see that all of the "natural" ways to obtain a point estimate from a Bayesian posterior can be interpreted in terms of Bayes rule's with different loss functions.

4.6 Bayesian hypothesis testing

The denominator that appears in Bayes' theorem is the Bayesian evidence and can be computed via

$$\mathcal{Z} = p(\mathbf{x}) = \int p(\mathbf{x} \mid \vec{\theta}) p(\vec{\theta}) \mathrm{d}\vec{\theta}.$$

When writing down Bayes' theorem we suppressed the fact that all of the quantities were conditioned on the particular model we were assuming for the data generating process. Explicitly reintroducing the dependence on the model, M, we have

$$p(\vec{\theta}|\mathbf{x}, M) = \frac{p(\mathbf{x}|\vec{\theta}, M) p(\vec{\theta}|M)}{p(\mathbf{x}|\mathbf{M})}.$$

This makes it clear that the evidence, $p(\mathbf{x}|\mathbf{M})$, represents the *probability of seeing the model data under model M* and can be thought of as the likelihood for the model given the observed data. If we now have more than one model, M_1 and M_2 say, that we believe could describe the data, we can compute the **posterior odds ratio** for M_1 over M_2

$$O_{12} = \frac{p(\mathbf{x}|\mathbf{M_1})}{p(\mathbf{x}|\mathbf{M_2})} \frac{p(M_1)}{p(M_2)}.$$

The first term is called the **Bayes factor** and is the ratio of the model likelihoods. The second term is the **prior odds ratio**, which represents our prior belief about the relative probability of the two models. The posterior odds is the ratio of model probabilities based on the observed data and is the basis for Bayesian hypothesis testing. For $O_{12} \gg 1$ we favour model M_1, while for $O_{12} \ll 1$ we favour M_2.

In the case of a flat prior on models the prior odds ratio is just 1 and decisions are based on the Bayes factor. Kass and Rafferty (1995) described a 'rule of thumb' for interpreting Bayes' factors. This is summarised in Table 2. This Table can be used to interpret the results of Bayesian hypothesis tests. Alternatively, the distribution of the Bayes factor can be computed under the null hypothesis and used, in a frequentist way, to produce a mapping between p-values and Bayesian posterior odds ratios.

The models M_1 and M_2 need not be very different, but could, for example, represent different regions of the parameter space of a distribution, e.g., $M_1 : \theta \in \Theta_1$ versus $M_2 : \theta \in \Theta_2$. If the two hypotheses are both simple then the Bayes factor reduces to the likelihood ratio, which we saw was the optimal test statistic in the frequentist hypothesis testing context.

Computation of the Bayesian evidence is challenging. Most sampling algorithms that return independent samples from the posterior ignore the evidence as it is just a normalisation constant. The evidence can be written as an integral over the posterior which can be

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

Table 2: Table for intepretation of Bayes' factors, as presented in Kass and Rafferty (1995).

approximated by a sum over samples

$$\frac{1}{\mathcal{Z}} = \int \frac{1}{p(\mathbf{x} \mid \vec{\theta})} \; \frac{p(\mathbf{x} \mid \vec{\theta})p(\vec{\theta})}{\mathcal{Z}} \mathrm{d}\vec{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{1}{p(\mathbf{x} \mid \vec{\theta}_i)}.$$

In other words it is the harmonic mean of the likelihoods of the samples. This is an extremely unstable approximation, however, as this sum is dominated by points with small likelihoods, but these are precisely the regions where there will be fewer samples and hence larger Monte Carlo error. Other techniques, such as nested sampling, can be used to compute evidences more accurately and these will be discussed in the next chapter.

Example: Suppose we have a two dimensional Normal likelihood of the form

$$p(\mathbf{x}|\vec{\theta}) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + 2\frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right] \qquad (67)$$

and use priors for the parameters μ_1 and μ_2 of the form

$$p(\mu_1) = \frac{1}{\Sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_1^2}\mu_1^2\right], \qquad p(\mu_2) = \frac{1}{\Sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_2^2}\mu_2^2\right]. \qquad (68)$$

We are interested in comparing the two models

$$M_1 : \mu_2 = 0, \qquad M_2 : \mu_2 \in (-\infty, \infty).$$

The evidence for M_1 can be computed as

$$\mathcal{Z}_1 = \frac{1}{2\pi\sigma_2}\sqrt{\frac{1-\rho^2}{\sigma_1^2+\Sigma_1^2}} \exp\left[-\frac{x_2^2(\sigma_1^2-(1-\rho^2)\Sigma_1^2) + 2\rho x_1 x_2 \sigma_1\sigma_2 + \sigma_2^2 x_1^2}{2\sigma_2^2(\sigma_1^2+\Sigma_1^2)}\right]$$

and for M_2 it is

$$\mathcal{Z}_2 = \frac{1}{2\pi}\sqrt{\frac{1-\rho^2}{\sigma_1^2(\sigma_2^2+\Sigma_2^2)+\Sigma_1^2(\sigma_2^2+(1-\rho^2)\Sigma_2^2)}} \times$$
$$\times \exp\left[-\frac{x_2^2((1-\rho^2)\Sigma_1^2+\sigma_1^2) + 2\rho x_1 x_2\sigma_1\sigma_2 + x_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)}{2\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2) + 2\sigma_1^2(\sigma_2^2+\Sigma_2^2)}\right] \qquad (69)$$

which gives the posterior odds ratio in favour of M_2, for equal prior odds (which is just the Bayes factor)

$$\mathcal{O}_{21} = \frac{\mathcal{Z}_2}{\mathcal{Z}_1} = \sigma_2\sqrt{\frac{\Sigma_1^2+\sigma_1^2}{\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)+\sigma_1^2(\Sigma_2^2+\sigma_2^2)}} \times$$
$$\times \exp\left[\frac{\Sigma_2^2(x_2((1-\rho^2)\Sigma_1^2+\sigma_1^2)+\rho x_1\sigma_1\sigma_2)^2}{2(\Sigma_1^2+\sigma_1^2)\sigma_2^2(\sigma_1^2(\Sigma_2^2+\sigma_2^2)+\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2))}\right]. \qquad (70)$$

This is difficult to interpret, but if we now assume that $\Sigma_1^2 \gg \sigma_1^2$, i.e., that the prior in μ_1 is much broader than the typical measurement uncertainty, the odds ratio simplifies to

$$\mathcal{O}_{21} \approx \sigma_2 \sqrt{\frac{1}{(1-\rho^2)\Sigma_2^2 + \sigma_2^2}} \exp\left[\frac{(1-\rho^2)x_2^2}{2\sigma_2^2}\right]$$

We see that there is a competition between the size of the additional variable dimension (characterised by Σ_2) in the first term and the weight of evidence for the additional effect in the data (characterised by the second term). Only if the addition of the extra dimension significantly improves the fit to the data (characterised by x_2 which is effectively the peak of the posterior in μ_2 when that parameter is allowed to vary) should the more complex model be favoured. If the fit does not improve, then the addition of the extra dimension is penalised by the first term and so the more complex model should not be preferred. It is often said that Bayesian posterior odds ratios automatically encode the notion of “Occam's razor”, i.e., one should use the simplest model that adequately describes the data since adding extra degrees of freedom always improves a fit. This is the sense in which it is meant. Addition of extra dimensions typically includes a prior penalty, as we see here, which will lead to the disfavouring of an alternative model unless the likelihood shows a significantly great improvement when the extra degrees of freedom are included.

4.7 Predictive checking

In both a frequentist and a Bayesian context it is natural to ask whether the model is a good representation of the observed data. In the Bayesian context this is accomplished by using **predictive distributions**.

Definition: the **prior predictive distribution** is the probability distribution

$$p(\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{x}|\vec{\theta})p(\vec{\theta})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the assigned prior distribution and therefore represents our *a priori* belief about the distribution of data sets that would be observed. Similarly, we have the following

Definition: the **posterior predictive distribution** is the probability distribution

$$p(\mathbf{y}|\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{y}|\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the posterior probability based on the observed data $\mathbf{x}$ and is our expectation about the distribution of future data sets $\mathbf{y}$.

The posterior predictive distribution can be used to assess whether the observed data is unusual within the posterior distribution, which is an indicator about whether or not the model is a good fit. Based on the observed data $\mathbf{x}$ we generate a large number of new data sets $\{\mathbf{y}_1, \ldots \mathbf{y}_N\}$ that are similar to $\mathbf{x}$, i.e., they consist of the same number of observations. For each data set we compute a set of summary statistics, and hence obtain the distribution of the summary statistics over many realisations of the posterior predictive distribution. We can then assess the “p-value” of the observed data within these distributions. If it looks like an outlier in any one of these distributions this suggests the model is not a good fit. Suitable summary statistics could include the maximum, minimum, median, skewness, kurtosis etc. Ideally we choose summary statistics that are orthogonal to the model parameters to increase sensitivity, since we are using the data twice (once to compute the posterior and once to compare to the predictive distribution). Statistics that are effectively tuned to the observed data will tend to lie in the middle of the predictive distributions by construction, even if the model is poor. We will see an example of this in the next section.

4.8 Example: regression

To illustrate some of the ideas discussed above we will present a Bayesian analysis of a regression problem. We suppose that we have made measurements of a set of values, $\{y_i\}$, corresponding to sets of p known explanatory variables, $\{\mathbf{x}_i\}$, and we believe that these follow a linear relationship with equal variance normally distributed errors

$$y_i \sim N(\mathbf{x}_i^T \vec{\beta}, \sigma^2), \quad i = 1, \ldots, N.$$

We want to infer the parameters of the linear relationship, $\vec{\beta}$, and the unknown precision $\tau = 1/\sigma^2$. We use a Bayesian framework and so must write down prior distributions on these parameters. We can assume a separable prior

$$p(\vec{\beta}, \tau) = p(\tau) \prod_{i=1}^{p} p(\beta_j)$$

and take Normal priors for the β_j's and a Gamma prior for τ as these are conjugate priors in the Normal-Gamma model

$$\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j}), \quad \tau \sim \text{Gamma}(a, b).$$

In the absence of prior information it is reasonable to set $\mu_{\beta_j} = 0$. Inferred values of the coefficients that are non-zero then provide evidence for the existence of a relationship between the observed data and those explanatory variables. Setting σ_j^2 to a large value, say 10^4, indicates large uncertainty in the parameter values and avoids strong prior dependence in the results. For the prior on τ, it is usual to take small values of a and b, for example $a = b = 0.1$ or $a = b = 0.01$. However, such priors lead to a preferred value (i.e., a peak) in the prior and so the use of such priors is somewhat controversial.

To illustrate fitting such a model, we can use a standard data set, the MTCARS data set, which is available in the R statistical software package and may also be found online. The data set contains observations, y_i, of the miles driven per gallon in the i'th of 32 different models of car, with explanatory variables x_{i1}, the rear axle ratio, x_{i2}, the weight of the i'th car and x_{i3}, the time to drive 0.25 miles from rest. We fit the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1/\tau), \quad i = 1, \ldots 32,$$

with $\beta_j \sim N(0, 1000)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$. We can use statistical software (in this case R) to generate samples from the posterior. Techniques for doing this will be discussed in the next chapter, and in the associated practical. using these samples we can obtain a posterior mean and 95% symmetric credible interval for each parameter. These can be compared to the frequentist estimates of the same parameters and the frequentist 95% confidence interval (see problem sheet 1). This comparison is in Table 3.

	Bayesian results		Frequentist results	
Parameter	Posterior mean	95% credible interval	MLE	95% confidence interval
β_0	10.369	[-5.098,36.349]	11.395	[-5.134,27.922]
β_1	1.777	[-0.721,4.166]	1.750	[-0.857,4.169]
β_2	-4.335	[-5.702,-2.995]	-4.347	[-5.787,-3.009]
β_3	0.968	[0.449,1.493]	0.946	[0.410,1.482]
σ^2	6.978	[4.160,11.729]	6.554	—

Table 3: Comparison between Bayesian and frequentist estimates of the linear model fit to the MTCARS data set.

The results of the Bayesian fit are quite consistent between the two approaches, although there are some differences and the interpretation of the results is different. We now want to assess the quality of the results. In a frequentist setting, assessment of the quality of a linear model fit is done through the production of *studentised residuals* and *Q-Q plots*. A studentised residual is

$$\hat{\epsilon}_i = \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\beta}$ are the estimated parameters, $\hat{\sigma}$ is the esitmated standard deviaiton and h_{ii} is the i'th diagonal element of the matrix $H = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$. These quantities follow a student-t distribution which is why they are called studentised residuals. A $Q - Q$ plot is a plot of the distribution of these values against the theoretical distribution, which should be approximately a straight line if the model is a good description of the data.

We can construct analogous quantities in the Bayesian case, but now the parameters are described by distributions rather than point estimates. A point estimate can be constructed in a number of different ways — using posterior mean values, using a single draw from the posterior, or averaging over the full posterior. The latter approach involves computing the studentised residual for a large number of draws from the posterior and averaging them, and is called the *posterior mean of the residual*. Studentised residuals are plotted in various ways in Figure 5.

We can also produce posterior predictive checks as described in section 4.7. We compute realisations of similar data sets and estimate the distribution of various summary statistics which we then compare to the values in the observed data sets. In this case we compute the distributions of the minimum, maximum, median and skewness in repeated data sets. These are shown in Figure 6, along with the values in the observed data set. We see that the observed values lie within the distributions in all cases, except for skewness. Seeing that the observed data lies in the tail of the distribution may indicate a failure of the model. In this case we might want to try varying the assumption of normally distributed errors and homoskedacity (equal error variance).

The issue with the posterior predictive checks could indicate a failure of the model, or the influence of an outlying data point. One way to tackle this is to modify the model so that the distribution of the errors ϵ_i is no longer assumed to be normal. The most common approach is to replace the normal distribution by a t_ν-distribution, as these have heavier tails. This is referred to as **robust regression**. The degrees of freedom, ν, in the t_ν-distribution can be fixed to some reasonable value, or allowed to vary in a hierarchical model (see next section). In that case the prior on ν is usually taken to be a Gamma distribution, $\nu \sim \text{Gamma}(c, d)$.

For the MTCARS dataset we try this, using prior values $c = d = 0.1$, and then look at the

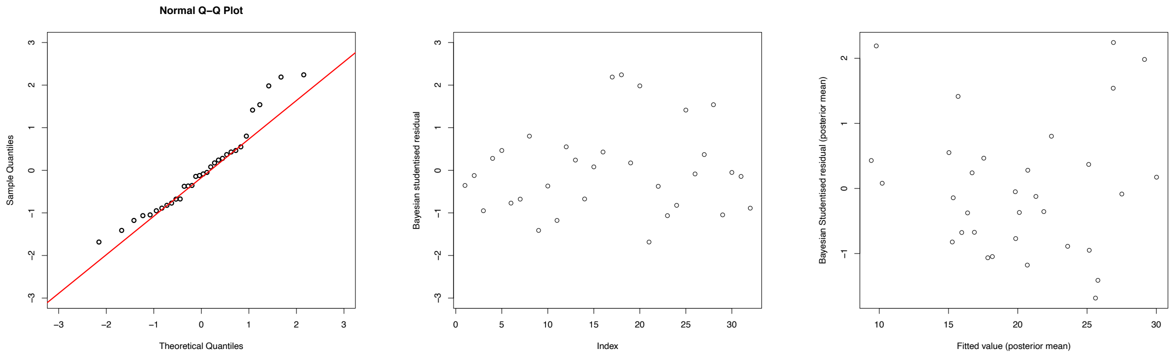

Figure 5: $Q-Q$ plot of the studentised residuals (left), studentised residual versus index of data point (middle) and studentised residual versus posterior mean of the predicted value, $\hat{y}_i$, for the Bayesian fit to the MTCARS data set. We look for the left hand plot to be on the diagonal line, for the middle and right hand plots we want the values to be randomly distributed (i.e., no trend with the x value) and in the range from minus a few to plus a few. These constraints are all satisfied here and so we see no cause for concern.

posterior predictive distribution again. The results for the skewness are shown in Figure 7. We that robustifying regression can help to improve the model fit in this case. The observed dat moves from lying at the 99.6% point of the distribution to lying at the 96.3%. So, it is still something of an outlier but it is not so much a cause for concern. It is perhaps not surprising that the use of robust regression only helped a small amount in this case, since we are trying to compensate for non-zero skew in the data and the t-distribution is also a symmetric distribution.

4.9 Hierarchical models

In many contexts, for example the observation of mergers of compact binary coalescences through gravitational wave observations, the likelihood describes the observation of a single event, and the prior describes the distribution of parameter values in the population from which the events are drawn. Often the parameters of the population prior are not themselves known but are of interest. For example, we do not know the distribution of masses of black holes in binaries and would like to learn about this from observations of the gravitational wave sources. This leads to the notion of a **hierarchical model**, in which the likelihood for data depends on parameters for which we write down a prior that in turn depends on unknown parameters (usually termed **hyperparameters**), for which we write down another prior (the **hyperprior**).

This hierarchy can be continued to more and more levels, but such models increase rapidly in complexity. Inference on complex hierarchical models can be simplified by imposing a *conditional independence* structure in the models, e.g., $p(x, y, z) = p(x|z)p(y|z)p(z)$. Conditional dependence structures can be compactly represented using *graphical models*. These are directed acyclic graphs that indicate dependencies between various components of the model. It is important that the graph has no cycles as only then can the joint probability be factorised. An example of a graphical model is shown in Figure 8. This model represents the following conditional dependence structure

$$p(p, q, r, s, t, u, v, w, x, y, z) = p(x|y, z)p(y|u, w)p(w|v)p(u)p(v)p(z|r)p(r|p, q)p(p)p(q) \quad (71)$$

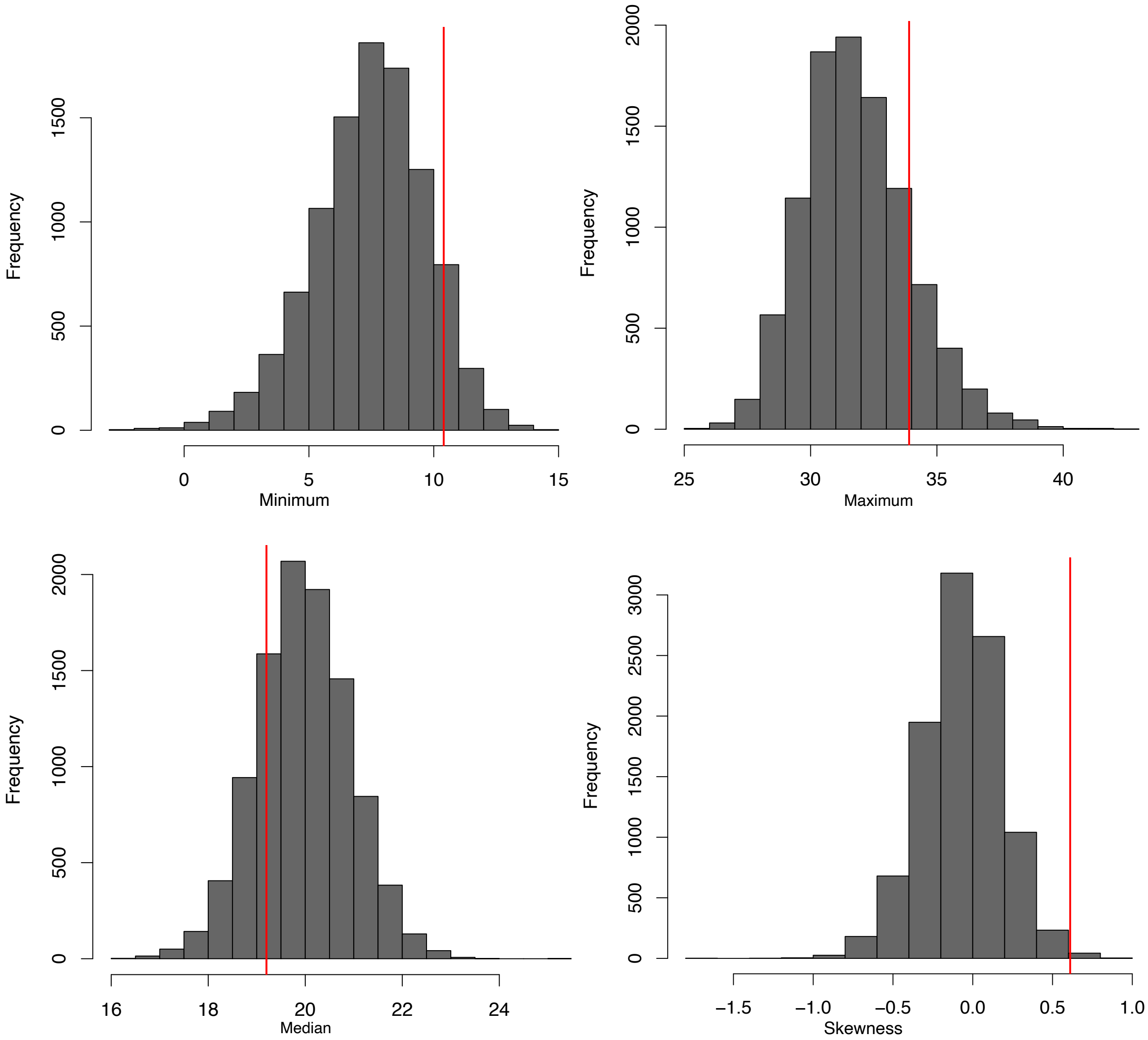

Figure 6: Predictive distributions for the maximum (top left), minimum (top right), median (bottom right) and skewness (bottom right) in replicated data sets of size 32, based on the posterior distribution from the MTCARS data set. The vertical red lines indicate the values in the data set form which the posterior was obtained. We see that this lies in the middle of the distribution in all cases, except skewness, in which it lies in the tail, which might indicate a failure to properly fit the data.

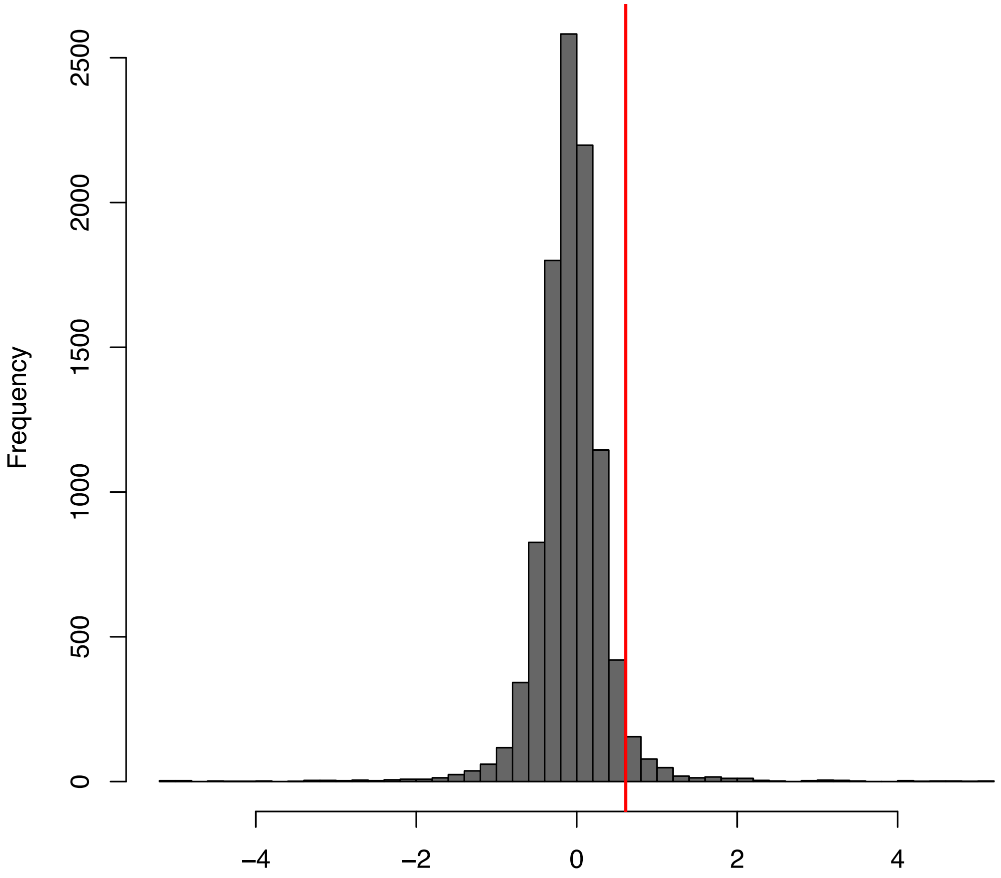

Figure 7: Posterior predictive distribution of skewness for the robustified regression model. The observed value of the skewness is indicated by a vertical red line as before.

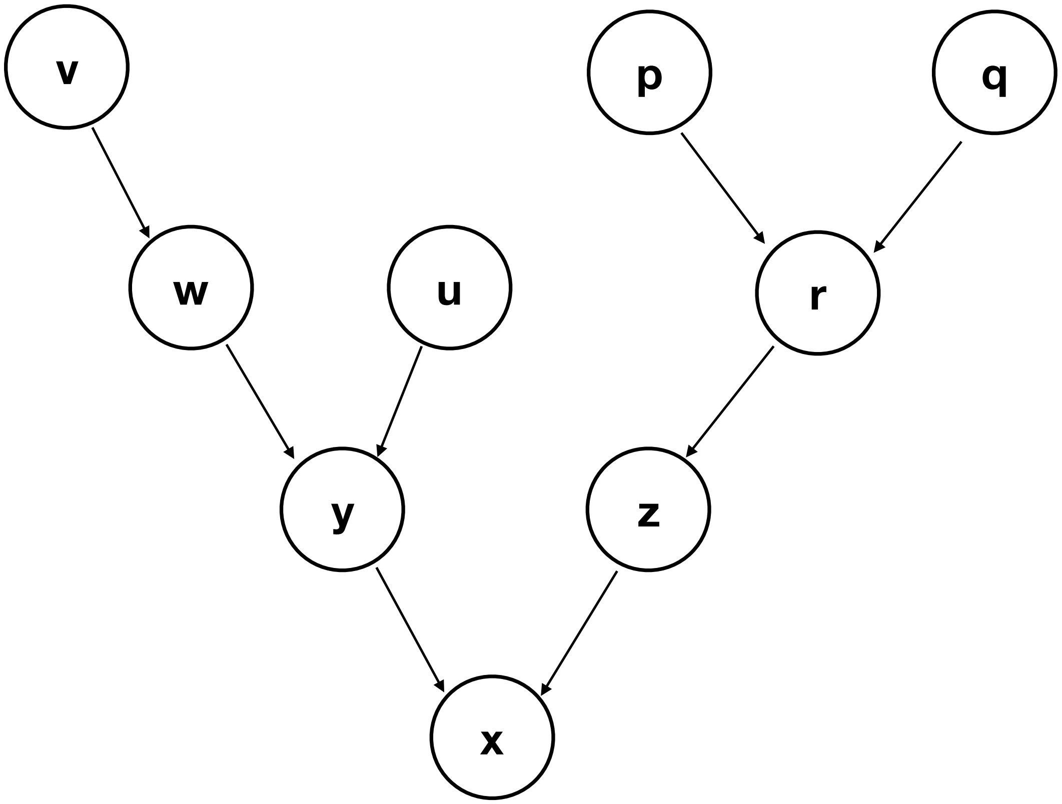

Figure 8: Illustration of a Bayesian graphical model. This is an acyclic directed graph that indicates conditional dependencies in complex Bayesian hierarchical models.

4.9.1 Selection effects

One thing that is important to account for in hierarchical modelling are selection effects. The decision about whether or not to include an event in a catalogue used for inference is based on whether or not the event is "detected", i.e., whether or not the observed data passes some pre-determined threshold criterion for inclusion. This is usually a property of the data only. Selection effects can be included by modifying the likelihood so that it represents the likelihood of "detected" data sets. If the un-corrected likelihood is $p(\mathbf{x}|\vec{\theta})$ then the likelihood for observed events is just

$$p(\mathbf{x}|\vec{\theta}, \text{obs}) = \frac{1}{p_s(\vec{\theta})} p(\mathbf{x}|\vec{\theta}), \quad \text{where } p_s(\vec{\theta}) = \int_{\mathbf{x}>\text{threshold}} p(\mathbf{x}|\vec{\theta}) \mathrm{d}\mathbf{x}.$$

The integral is over all data sets that would have been considered as "detections", i.e., passing the threshold for inclusion in inference. What we have done here is renormalise the likelihood so that it integrates to 1 over all above threshold data sets. Since the partition of the data into observed and unobserved is a property of $\mathbf{x}$ only, the relative probabilities of different above threshold data sets must be in proportion to their probabilities in the set of all data sets.

Usually, the likelihood will depend on parameters of the particular source, $\vec{\theta}$, that are themselves determined by the priors, which depends on the hyperparameters of the population, $\vec{\lambda}$. Then the likelihood for observed events, marginalised over the source parameters is simply

$$p(\mathbf{x}|\vec{\lambda},\text{obs}) = \frac{1}{p_s(\vec{\lambda})}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}, \quad \text{where } p_s(\vec{\lambda}) = \int_{\mathbf{x}>\text{threshold}}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x}.$$

Usually we are interested in the parameters of individual sources as well as the overall population parameters. The joint likelihood of $\mathbf{x}$ and $\vec{\theta}$, conditioned on detection, is

$$p(\mathbf{x},\vec{\theta}|\vec{\lambda},\text{obs}) = p(\mathbf{x}|\vec{\theta},\text{obs})p(\vec{\theta}|\vec{\lambda},\text{obs}).$$

The first term is Eq. (4.9.1), but for the source parameters $\vec{\theta}$

$$p(\mathbf{x}|\vec{\theta},\text{obs}) = \frac{p(\mathbf{x}|\vec{\theta})}{p(\text{obs}|\vec{\theta})}, \quad \text{where } p(\text{obs}|\vec{\theta}) = \int_{\mathbf{x}>\text{threshold}} p(\mathbf{x}|\vec{\theta})\mathrm{d}\mathbf{x}.$$

The second term is the prior on $\vec{\theta}$ *for events above threshold.* However, this prior is modified from $p(\vec{\theta}|\vec{\lambda})$ by the conditioning on detection, namely

$$p(\vec{\theta}|\vec{\lambda},\text{obs}) = \frac{p(\vec{\theta},\text{obs}|\vec{\lambda})}{p(\text{obs}|\vec{\lambda})} = \frac{p(\text{obs}|\vec{\theta},\vec{\lambda})p(\vec{\theta}|\vec{\lambda})}{p(\text{obs}|\vec{\lambda})} = \frac{p(\text{obs}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}.$$

Putting this together we see that the terms relating to selection on $\vec{\theta}$, $p(\text{obs}|\vec{\theta})$, cancel and the joint likelihood is

$$p(\mathbf{x},\vec{\theta}|\vec{\lambda},\text{obs}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}$$

giving a posterior on $\vec{\theta}$

$$p(\vec{\theta}|\mathbf{x},\vec{\lambda},\text{obs}) \propto p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})$$

which is unchanged from the posterior that would be written down if there is no selection. We see that the selection effects corrections do not change inference about the parameters of individual sources, only inference about the hyperparameters governing the population as a whole.

This approach implicitly assumes that the number of observed events contains no information about the unknown parameters. An alternative approach is to write down a joint likelihood for all events, both the N_{obs} events that are observed, $\{\mathbf{x}_i\}$, with parameters $\{\vec{\theta}_i\}$, and the N_{nobs} events that are unobserved, $\{\mathbf{x}_j\}$, with parameters $\{\vec{\theta}_j\}$. We model the number of events as a Poisson process with overall rate $N(\vec{\lambda})$, and rate density $\mathrm{d}N/\mathrm{d}\vec{\theta}$. The joint likelihood is

$$p\left(\left\{\vec{\theta}_i\right\},\left\{\vec{\theta}_j\right\},\{\mathbf{x}_i\},\{\mathbf{x}_j\}\mid\vec{\lambda}\right) \propto \left[\prod_{i=1}^{N\text{obs}} p\left(\mathbf{x}_i\mid\vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right]\times$$
$$\times\left[\prod_{j=1}^{N_{\text{nobs}}} p\left(\mathbf{x}_j\mid\vec{\theta}_j\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_j}\left(\vec{\lambda}\right)\right]\exp\left[-N\left(\vec{\lambda}\right)\right] \tag{72}$$

We can marginalise over the unobserved data to obtain

$$p\left(\left\{\vec{\theta}_i\right\},\{\mathbf{x}_i\}\mid\vec{\lambda}\right)\propto\quad\left[\prod_{i=1}^{N_{\text{obs}}}p\left(\mathbf{x}_i\mid\vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right]\frac{N_{\text{ndet}}^{N_{\text{nobs}}}\left(\vec{\lambda}\right)}{N_{\text{nobs}}!}\exp\left[-N\left(\vec{\lambda}\right)\right]\tag{73}$$

where

$$N_{\text{ndet}}\left(\vec{\lambda}\right)\equiv\int_{\{\mathbf{x}<\text{threshold}\}}\mathrm{d}\mathbf{x}\,\mathrm{d}\vec{\theta}\,p\left(\mathbf{x}\mid\vec{\theta}\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}}\left(\vec{\lambda}\right).\tag{74}$$

We can then marginalise over the unknown number of unobserved events to obtain

$$p\left(\left\{\vec{\theta}_i\right\},\{\mathbf{x}_i\}\mid\vec{\lambda}\right)\propto\left[\prod_{i=1}^{N_{\text{obs}}}p\left(\mathbf{x}_i\mid\vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right]\exp\left[-N_{\text{det}}\left(\vec{\lambda}\right)\right].\tag{75}$$

We can now introduce the overall rate in the Unvierse, N, by writing $\mathrm{d}N/\mathrm{d}\vec{\theta}=Np(\vec{\theta}|\vec{\lambda})$. Then

$$N_{\text{det}}(\vec{\lambda})=N\int_{\mathbf{x}>\text{threshold}}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x}=Np_s(\vec{\lambda}).\tag{76}$$

Setting a scale-invariant prior on N (which states that the number of detected events does not convey information about the unknown parameters of the population), $p(N)\propto 1/N$ we can marginalise N out of the likelihood and recover Eq. (4.9.1).

4.9.2 Examples of hierarchical models

We finish this section with two examples of Bayesian hierarchical models.

Example 1: Salmon fishery In a given year, several fish hatcheries located along rivers in Washington state, USA raise coho salmon from eggs to a juvenile stage. Each hatchery releases a batch of juvenile fish into the rivers. The fish then travel to the ocean and some of them return to the hatchery 3 years later. The probability that a juvenile salmon returns varies between hatcheries due to different hatchery practices and river conditions at the point of release. We construct a hierarchical model for this as follows

- Suppose there are J fisheries and n_j salmon observed at fishery j.
- The data for an individual observation, x_{ji}, of the i'th salmon at fishery j is Bernoulli (salmon returned or did not return), with parameter p_j, where j labels the fishery. The data for the total number of returning salmon at site j, x_j, is Binomial with parameters (n_j,p_j).
- We assume that the p_j's are drawn from some common global distribution and use the conjugate prior of Beta(a,b).
- The parameters a and b are not known and fixed as in the usual case, but these are unknown quantities of interest as they characterise the variability in the population. These are the hyperparameters of the prior on p_j.
- We define a suitable hyperprior $p(a,b)$ on the hyperparameters, for example a Gamma prior.

- The joint posterior on the set $(\{p_j\}, a, b)$ is

$$p(\{p_j\}, a, b|\mathbf{x}) \propto p(\mathbf{x}|\{p_j\}) \left[\prod_{j=1}^{J} p(p_j|a, b)\right] p(a, b).$$

  Note that the hyperprior on the hyperparameters appears only once as these parameters are common to all of the individual observations of fisheries.

- The marginal distribution on the hyperparameters (a, b) can be found by marginalising over the $\{p_j\}$'s

$$p(a, b|\mathbf{x}) \propto p(a, b) \prod_{j=1}^{J} \frac{B\left(a + x_j, b + n_j - x_j\right)}{B(a, b)}.$$

- Marginals on individual p_j's can be found in a similar way.

Example 2: Gravitational wave cosmology In August 2017 the LIGO/Virgo gravitational wave detectors observed gravitational waves from the inspiral and merger of a binary neutron star for the first time, GW170817. There was both a short gamma ray burst and a kilonova associated with this event, which allowed the unique identification of the host galaxy, NGC 4993, and hence the recessional velocity (redshift) of the host. The gravitational waves provide a measurement of the luminosity distance of the source. The rate of expansion of the Universe as a function of distance is a key observable for constraining cosmological parameters. The relationship is linear at low distances and the constant of proportionality is called the *Hubble constant*,

$$v = cz = H_0 d,$$

where v is the recessional velocity due to the expansion of the Universe, z is the corresponding redshift, H_0 is the Hubble constant and d is the luminosity distance. At low distance/redshift, the *peculiar velocity* of individual galaxies, relative to the overall expansion of the Universe (the "Hubble flow") is significant and so the observed recessional velocity, v_r, must be corrected by writing $v_r = H_0 d + v_p$. Observations of galaxies provide an estimate of the smoothed peculiar velocity field, $\langle v_p \rangle$. We are interested in inferring the value of the Hubble constant and build a hierarchical model as follows.

- The observed gravitational wave data, x_{GW}, depends on the waveform of the source, which in turn depends on the source parameters. Most of these are not of interest, denoted $\vec{\lambda}$, and so we can marginalise them out, but we treat distance d and inclination, ι, separately

$$p(x_{\mathrm{GW}} \mid d, \cos\iota) = \int p(x_{\mathrm{GW}} \mid d, \cos\iota, \vec{\lambda})\, p(\vec{\lambda}) \mathrm{d}\vec{\lambda}. \qquad (77)$$

- The measured recessional velocity, v_r, depends on the true recessional velocity, which depends on the peculiar velocity, v_p, and the Hubble redshift, $H_0 d$. Representing the electromagnetic measurement uncertainty as a Normal distribution we have

$$p\left(v_r \mid d, v_p, H_0\right) = N\left[v_p + H_0 d, \sigma_{v_r}^2\right]\left(v_r\right) \qquad (78)$$

- The measured smoothed peculiar velocity field at the location of the host galaxy depends on the true peculiar velocity there (and perhaps also on other quantities, but we suppress other dependencies here)

$$p\left(\langle v_p \rangle \mid v_p\right) = N\left[v_p, \sigma^2_{v_p}\right]\left(\langle v_p \rangle\right). \tag{79}$$

- The combined likelihood for the observations of $x_{\rm GW}$, $\langle v_p \rangle$ and v_r is

$$p(x_{\rm GW}, v_r, \langle v_p \rangle \mid d, \cos\iota, v_p, H_0) = \frac{1}{\mathcal{N}_s(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0)\, p(\langle v_p \rangle \mid v_p). \tag{80}$$

Here the factor $\mathcal{N}_s(H_0)$ is the selection effects factor discussed earlier, which corrects for the fact that we only analyse events that exceed some threshold in the gravitational wave detector

$$\mathcal{N}_{\rm s}(H_0) = \int_{\rm detectable} {\rm d}\vec{\lambda}\, {\rm d}d\, {\rm d}v_p\, {\rm d}\cos\iota\, {\rm d}x_{\rm GW}\, {\rm d}v_r\, {\rm d}\langle v_p \rangle \times \Big[p(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda})\, p(v_r \mid d, v_p, H_0) \times p(\langle v_p \rangle \mid v_p)\, p(\vec{\lambda})\, p(d)\, p(v_p)\, p(\cos\iota)\Big], \tag{81}$$

At the time of GW170817 the horizon for detection of binary neutron stars by the LIGO/Virgo detectors was much smaller ($\sim$ 100Mpc) than the distance to which the kilonova radiation could have been confidently observed ($\sim$ 400Mpc). This means that gravitational wave selection effects were dominant. As these depend directly on the luminosity distance, the dependence on H_0 is a higher order correction and so the selection function was approximately independent of H_0. A correct treatment of election effects will become increasingly important as the LIGO horizon increases in the future.

- We define priors on H_0, d, v_p and $\cos\iota$. These are independent and so we write down a product prior

$$p(d, \cos\iota, v_p, H_0) = p(d)p(\cos\iota)p(v_p)p(H_0).$$

We use flat priors on $\cos\iota$ and v_p, a volumetric prior on d, $p(d) \propto {\rm d}V_c/{\rm d}d$, where V_c is the comoving volume. We leave $p(H_0)$ unspecified, but note that the analysis in Abbott et al. (2017) used a scale-invariant prior $p(H_0) \propto 1/H_0$.

- We have now fully specified the hierarchical model. A graphical representation of this model is given in Figure 9. The posterior can now be found as

$$p(H_0, d, \cos\iota, v_p \mid x_{\rm GW}, v_r, \langle v_p \rangle) \propto \frac{p(H_0)}{\mathcal{N}_{\rm s}(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \times p(\langle v_p \rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota), \tag{82}$$

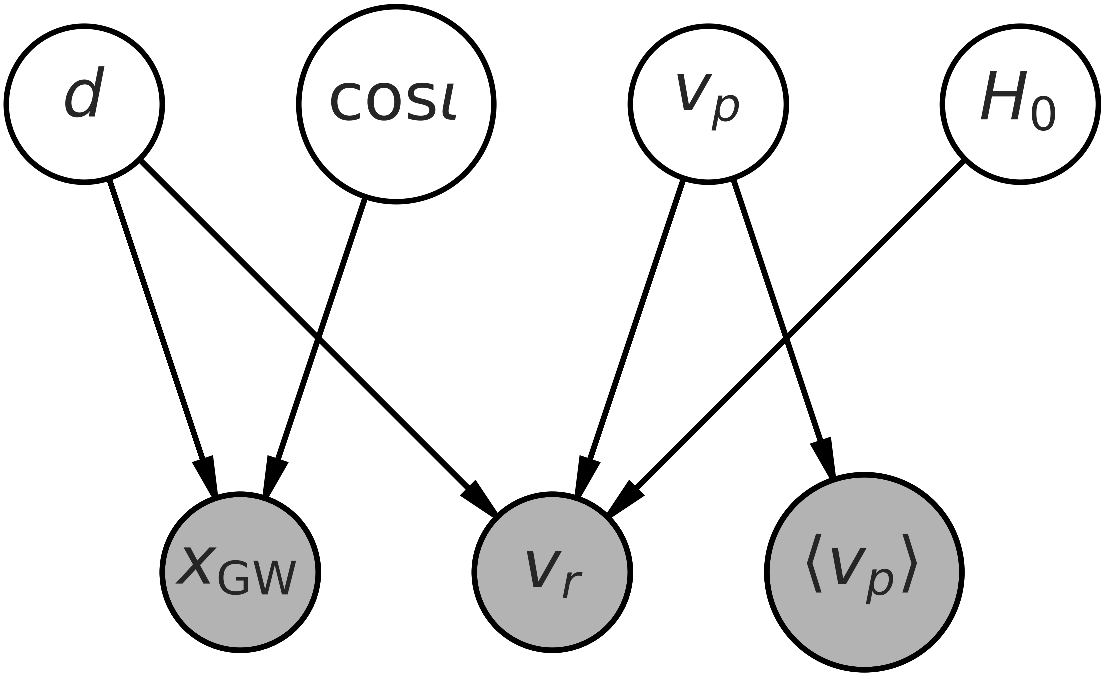

Figure 9: Graphical model for the Hubble constant measurement with gravitational wave observations of binary neutron stars. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).

- This posterior can be marginalised over d, $\cos\iota$ and v_p to give

$$\begin{aligned}p(H_0 \mid x_{\mathrm{GW}}, v_r, \langle v_p\rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} \int & \mathrm{d}d\, \mathrm{d}v_p\, \mathrm{d}\!\cos\iota \\ & \times p(x_{\mathrm{GW}} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \\ & \qquad \times p(\langle v_p\rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)\,. \quad (83)\end{aligned}$$

  This marginalised posterior is shown in Figure 10.

- If we make subsequent observations of binary neutron star mergers with counterparts, indexed by a superscript $i = 1, \ldots, N$, we can combine these

$$\begin{aligned}p(H_0 \mid \{x^i_{\mathrm{GW}}, v^i_r, \langle v_p\rangle^i\}) \propto \frac{p(H_0)}{\mathcal{N}^N_s(H_0)} \prod_{i=1}^{N} \Big[\int & \mathrm{d}d\, \mathrm{d}v_p\, \mathrm{d}\!\cos\iota \\ & \times p(x^i_{\mathrm{GW}} \mid d, \cos\iota)\, p(v^i_r \mid d, v_p, H_0) \\ & \qquad \times p(\langle v_p\rangle^i \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota) \Big]\,. \quad (84)\end{aligned}$$

  Note that, as in the previous example, the prior on the common hyperparameters, $p(H_0)$, occurs only once. The selection effect correction appears once for every observation.

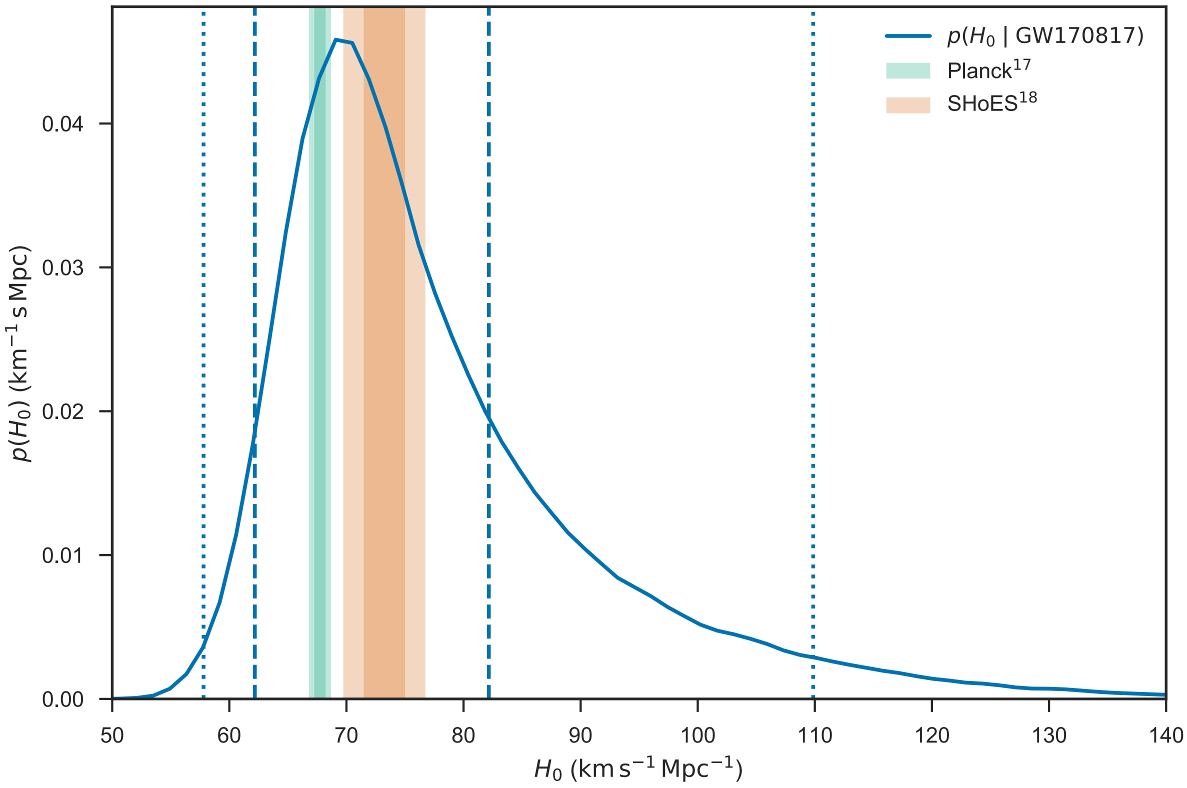

Figure 10: Posterior on the Hubble constant derived from GW170817. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).