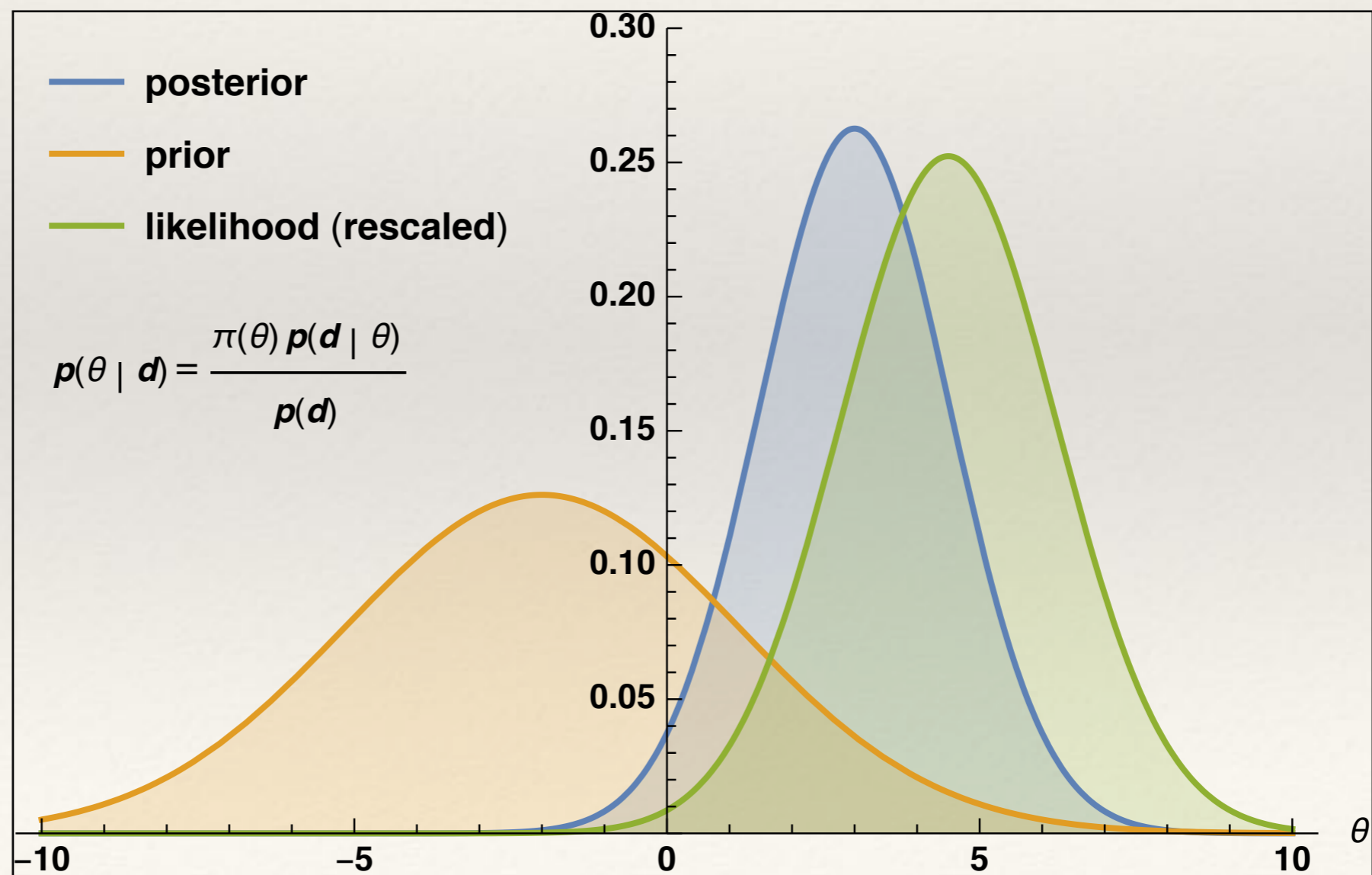


# Making sense of data: introduction to statistics for gravitational wave astronomy

## Lecture 5: Bayesian inference part II

*AEI IMPRS Lecture Course*

*Jonathan Gair jgair@aei.mpg.de*



---

# Bayesian hypothesis testing

---

- ❖ The denominator in Bayes' Theorem

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

- ❖ is the **Bayesian evidence**

$$p(\mathbf{x}|M) = \int p(\mathbf{x}|\vec{\theta}, M)p(\vec{\theta}|M) d\vec{\theta}$$

- ❖ Here we have explicitly introduced the model  $M$  to emphasises that the result depends on the model assumed. The evidence is the probability that the observed data would have been produced under the given model and so can be used for **model selection**.
- ❖ Models are compared using the **posterior odds ratio**

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

- ❖ The first term is the **Bayes Factor**. The second is the **prior odds ratio**.

---

# Bayesian hypothesis testing

---

- ❖ The interpretation of the posterior odds ratio is somewhat arbitrary, but Kass and Raftery (1995) suggested the following scale:

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 $M_1$

- ❖ Interpreting the Bayes factor as a ratio of probabilities, these thresholds corresponds to “p-values” of 0.25, 0.05, 0.007, but the interpretation is different.
- ❖ In practice, posterior odds ratios can also be used as a test statistic, with significance and power computed via simulation in the usual (frequentist) way.

---

# Bayesian hypothesis testing

---

- ❖ Computing Bayesian evidences is challenging. These can be estimated using the **harmonic mean of the likelihood** of samples from the posterior

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

- ❖ Necessarily, there are more posterior samples where the likelihood and hence posterior are higher.
- ❖ Regions where the likelihood is small are less well sampled and subject to more Monte Carlo error. This makes the above expression very unstable and potentially inaccurate.
- ❖ Other techniques, such as **nested sampling**, have been developed to overcome these problems and produce robust evidence estimates.



---

# Bayesian hypothesis testing

---

- ❖ **Example:** Normal models. Suppose we have a 2-dimensional likelihood

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

- ❖ and set priors of the form

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

- ❖ and we want to test the models

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

- ❖ The evidence for model 1 is

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

# Bayesian hypothesis testing

- ❖ The evidence for model 2 is

$$\mathcal{Z}_2 = \frac{1}{2\pi} \sqrt{\frac{1 - \rho^2}{\sigma_1^2(\sigma_2^2 + \Sigma_2^2) + \Sigma_1^2(\sigma_2^2 + (1 - \rho^2)\Sigma_2^2)}} \times$$

$$\times \exp \left[ -\frac{x_2^2((1 - \rho^2)\Sigma_1^2 + \sigma_1^2) + 2\rho x_1 x_2 \sigma_1 \sigma_2 + x_1^2((1 - \rho^2)\Sigma_2^2 + \sigma_2^2)}{2\Sigma_1^2((1 - \rho^2)\Sigma_2^2 + \sigma_2^2) + 2\sigma_1^2(\sigma_2^2 + \Sigma_2^2)} \right]$$

- ❖ giving a posterior odds ratio

$$\mathcal{O}_{21} = \frac{\mathcal{Z}_2}{\mathcal{Z}_1} = \sigma_2 \sqrt{\frac{\Sigma_1^2 + \sigma_1^2}{\Sigma_1^2((1 - \rho^2)\Sigma_2^2 + \sigma_2^2) + \sigma_1^2(\Sigma_2^2 + \sigma_2^2)}} \times$$

$$\times \exp \left[ \frac{\Sigma_2^2(x_2((1 - \rho^2)\Sigma_1^2 + \sigma_1^2) + \rho x_1 \sigma_1 \sigma_2)^2}{2(\Sigma_1^2 + \sigma_1^2)\sigma_2^2(\sigma_1^2(\Sigma_2^2 + \sigma_2^2) + \Sigma_1^2((1 - \rho^2)\Sigma_2^2 + \sigma_2^2))} \right]$$

- ❖ which can be simplified assuming  $\Sigma_1^2 \gg \sigma_1^2$

Size of extra dimension

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

Improvement in fit to data

- ❖ This can be interpreted as automatically implementing *Occam's Razor*.

---

# Predictive checking

---

- ❖ It is natural to want to test if the assumed model is a good fit to the data. In a Bayesian context this is achieved through **predictive checking**.

- ❖ The **prior predictive distribution** is defined by

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

- ❖ This is the distribution of observed data sets within the model assumed in the prior. If the observed data is not very consistent with this distribution, the prior parameters might need to be adjusted.

- ❖ The **posterior predictive distribution** is defined similarly

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

- ❖ This is the distribution of new datasets based on the best model fit to the data. The observed data should lie within the body of this distribution if the model is good.

---

# Example: linear model

---

- ❖ The predictive distribution can be used to compute the distribution of summary quantities. The value of those summary quantities in the observed data can then be compared to these distributions.

- ❖ Recall the linear model we fit to the mtcars data set

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

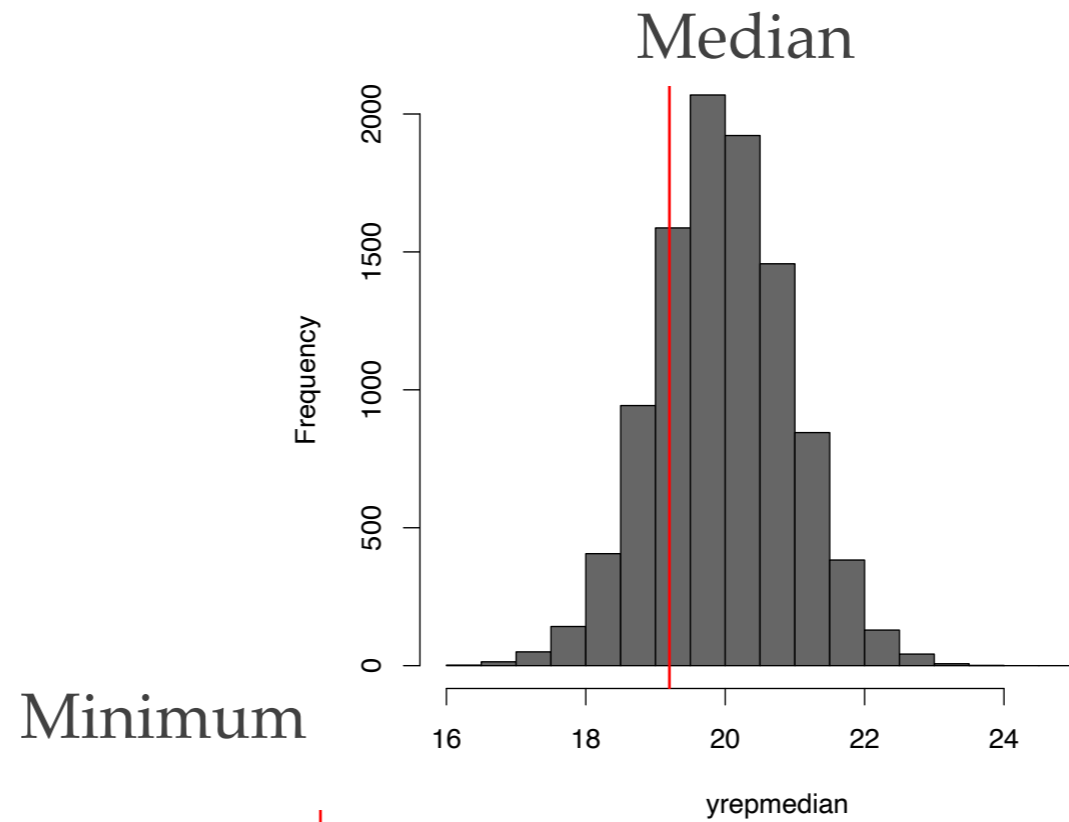
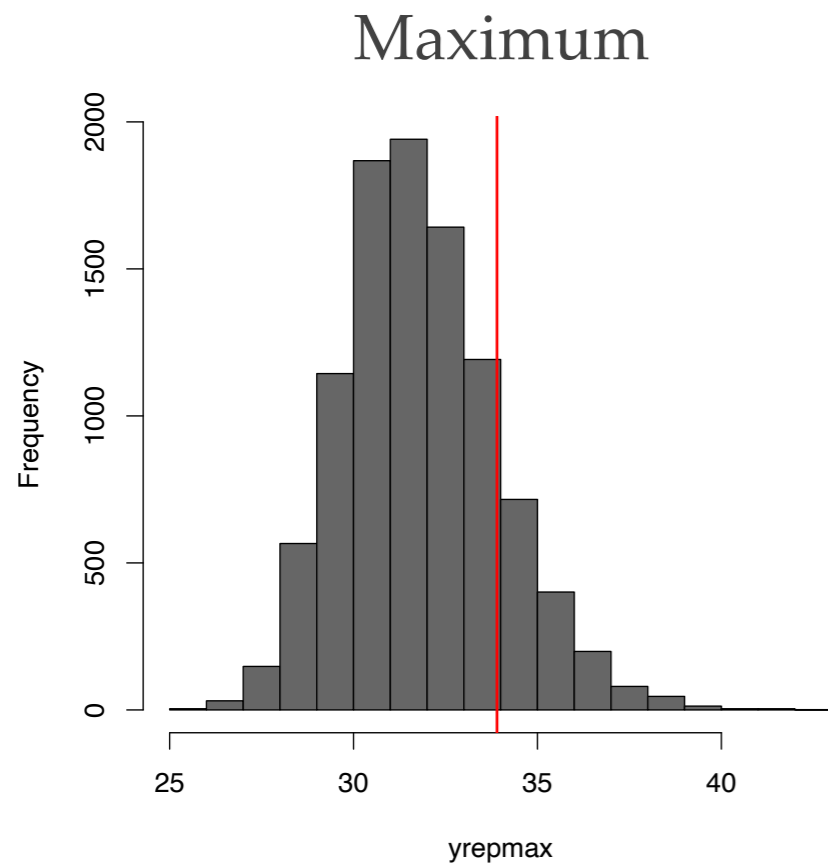
$$p(\vec{\beta}, \tau) = p(\tau) \prod_{j=1}^p p(\beta_j)$$

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

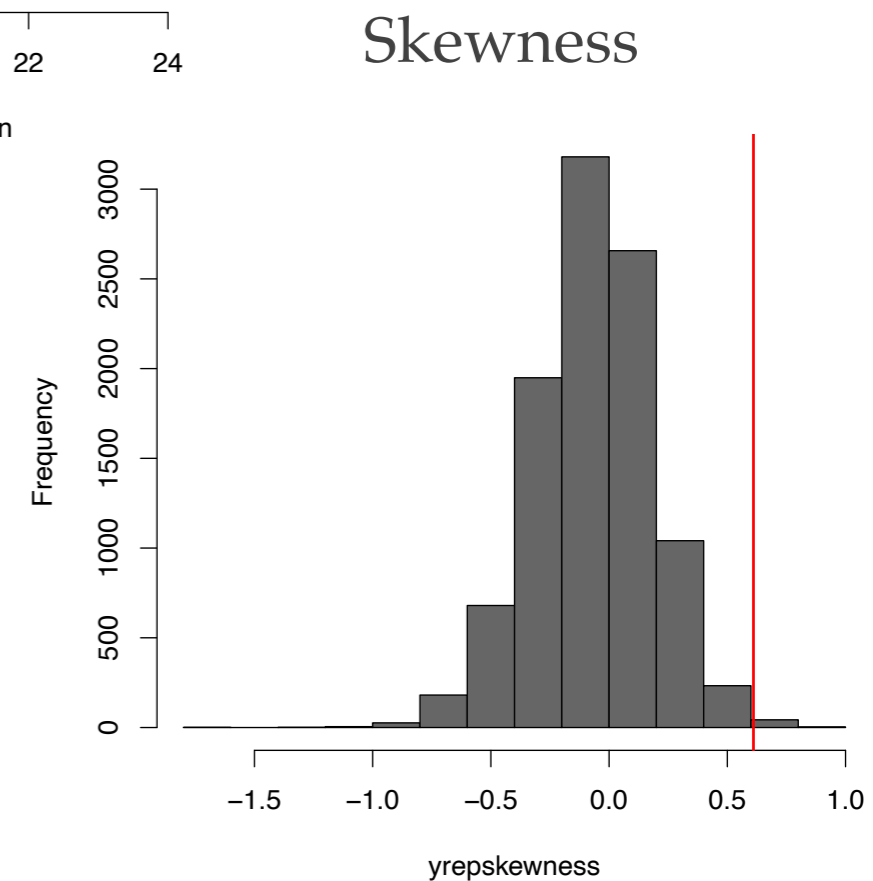
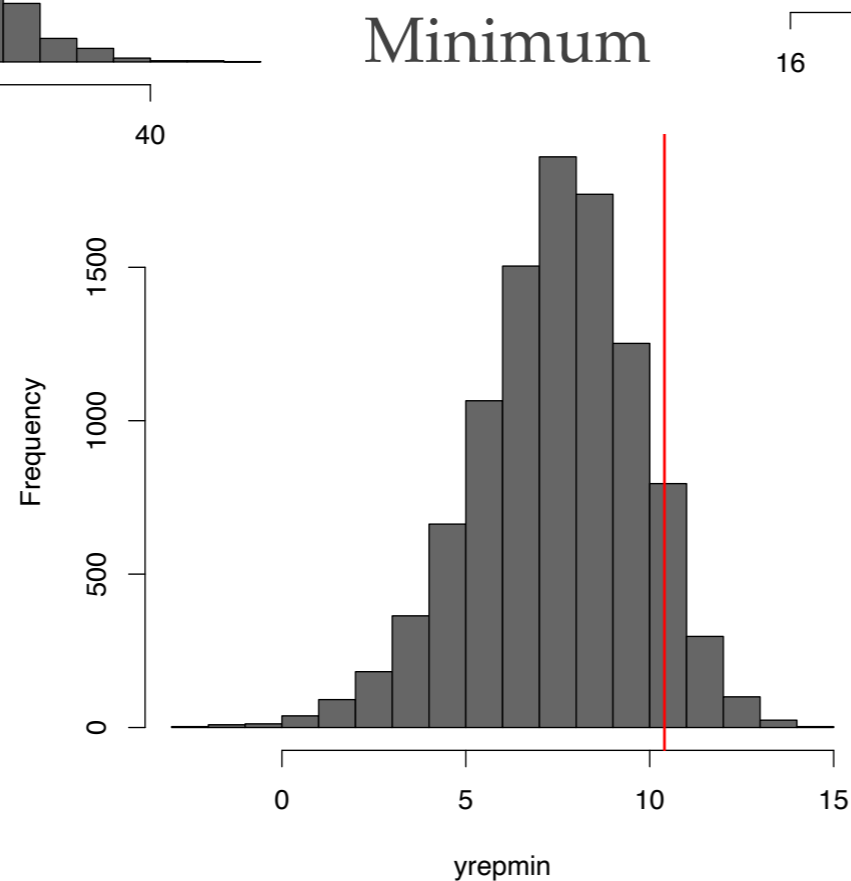
- ❖ We fit the model and compute the predictive distribution of the minimum, maximum, median and skewness of future samples of the same size.
- ❖ It is better to choose quantities that are somewhat “orthogonal” to what is adjusted to fit the data.



# Example: linear model



Skewness is a bit of an outlier - may indicate a model failure



---

# Robust regression

---

- ❖ If the posterior predictive distribution indicates some problems, or inspection of the data reveals the presence of some outliers, it is common practice to use **robust regression**.

- ❖ This involves replacing the Normal distribution by a  $t$ -distribution in the model

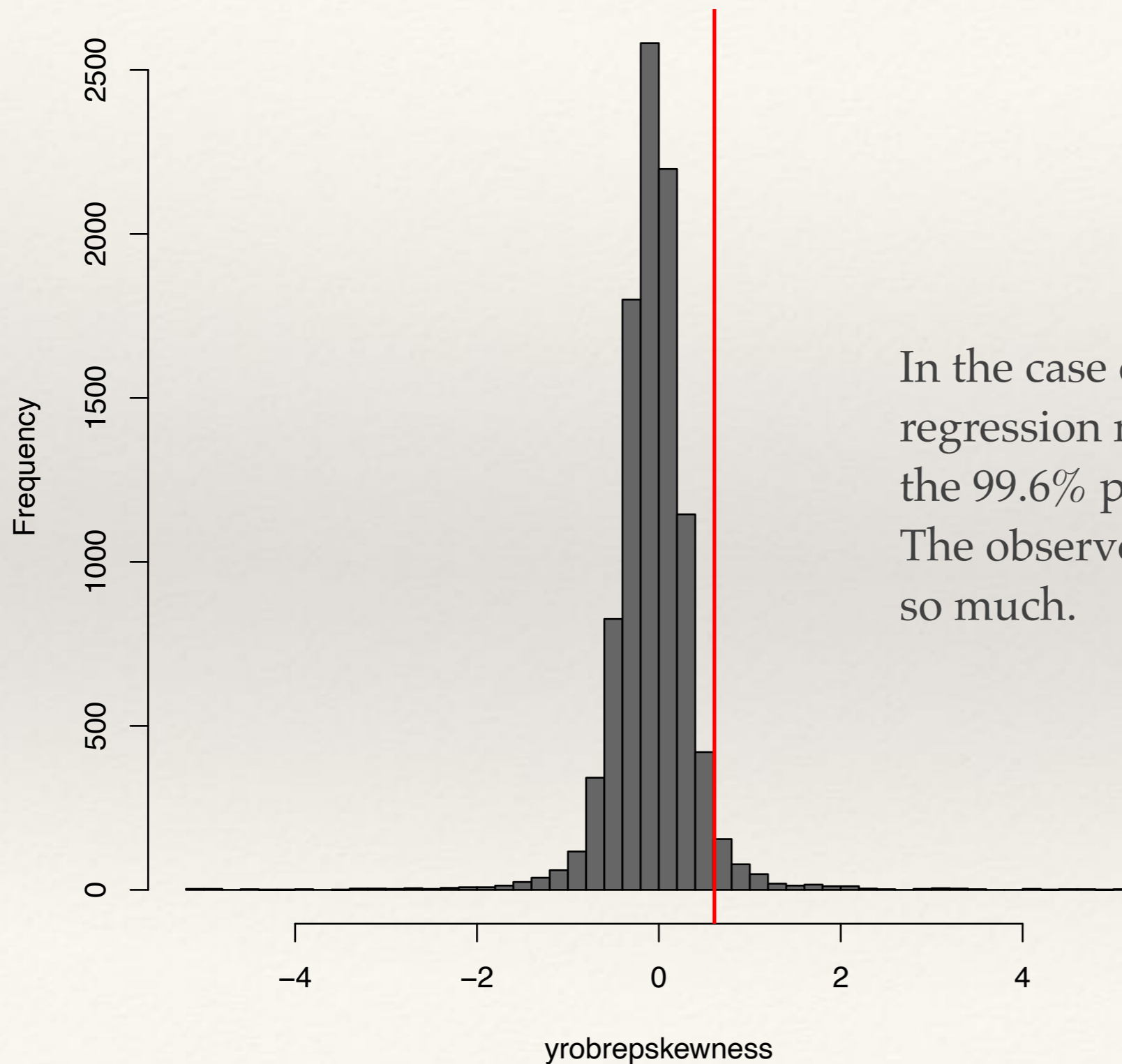
$$y_i \sim \mathbf{x}_i^T \beta + \sigma t_\nu$$

- ❖ The degrees of freedom can be fixed or treated as a parameter to be varied, in which case we require a prior. A Gamma prior is appropriate

$$\nu \sim \text{Gamma}(c, d)$$

- ❖ The  $t$ -distribution has heavier tails than the Normal distribution and so is better able to fit data that has outliers.
- ❖ This approach has been used in GW parameter estimation in (the **student-t likelihood** of Röver et al.). In that context it also arises from marginalisation over power spectral density uncertainty.

# Robust regression



In the case of the mtcars data set, robust regression moves the observed data from the 99.6% percentile to the 96.3% percentile. The observed data is still an outlier, but not so much.

---

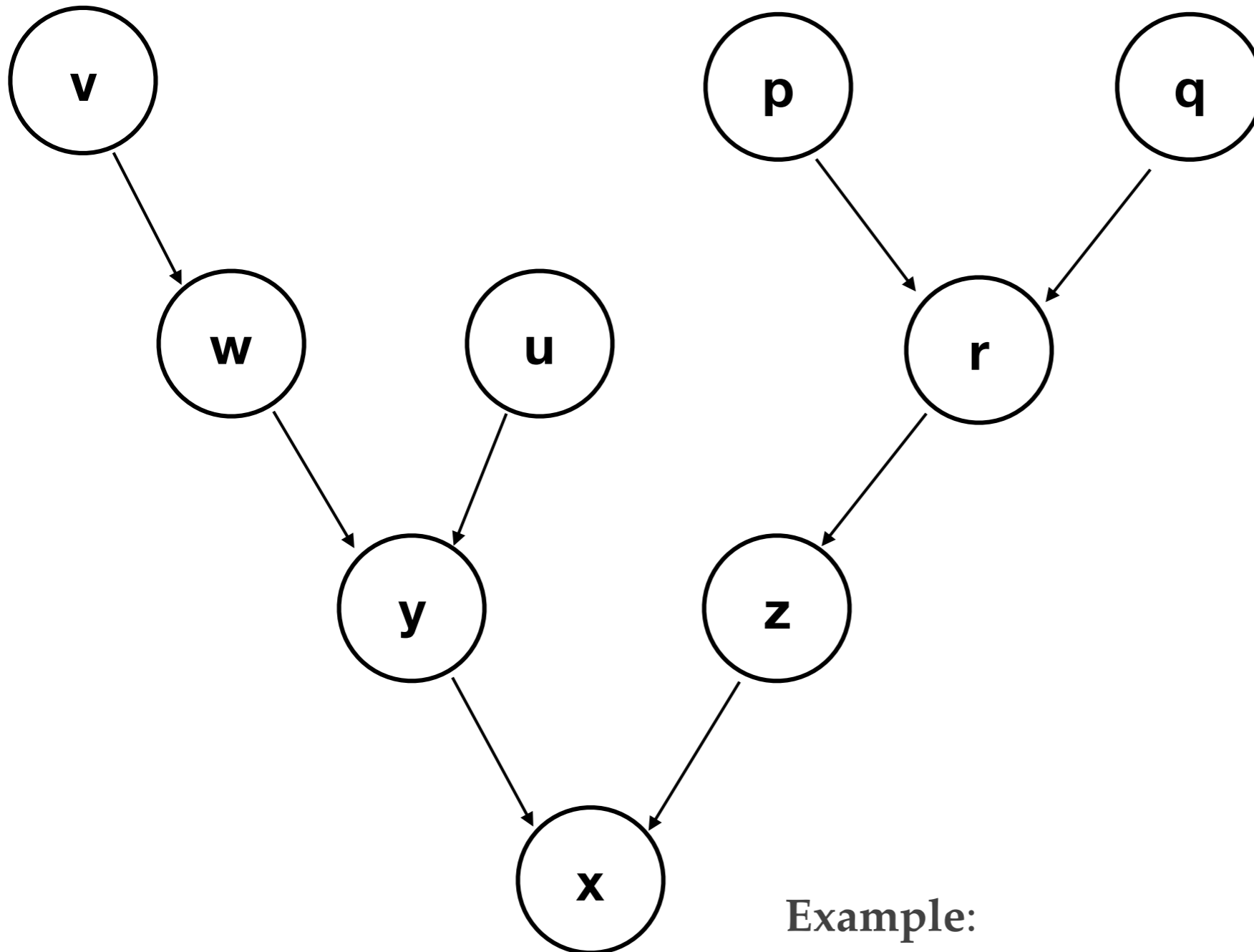
# Hierarchical Models

---

- ❖ Often the prior for a single data set represents a model for a population of events, e.g., compact binary coalescences.
- ❖ The parameters of that prior encode the details of the population and are also of interest. This leads to the notion of a **hierarchical model**.
- ❖ In hierarchical model, the parameters of the prior (termed **hyperparameters**) are regarded as random variables, on which a **hyperprior** is defined. This can be continued at infinitum - using another hyperprior on the hyperparameters of the first hyperprior etc.
- ❖ Such models can quickly get complicated as additional layers are included. They can be simplified by imposing a **conditional independence** structure.
- ❖ Hierarchical models can be most easily summarised using **graphical models**, which are *directed acyclic graphs* showing conditional dependencies.



# Graphical Model



Example:

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

$$p(x | y, z) p(y | u, w) p(z | r) p(w | v) p(r | p, q) p(v) p(u) p(p) p(q)$$

---

# Selection Effects

---

- ❖ No instrument is arbitrarily sensitive and therefore some types of source are easier to see than others. This is important to remember in hierarchical models for populations when we are combining only **detected** events.
- ❖ There are two ways to think about selection effects.
- ❖ One way is to acknowledge we only include “detected” events in the analysis and then write down a likelihood for detected events. This must integrate to 1 over all “detected” or “above threshold” data sets.

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

- ❖ This framework assumes a priori that the number of detected events contains no information about the parameters of interest.

---

# Selection Effects

---

- ❖ Alternatively we write down the likelihood for all events, both **detected** events (indexed by  $i$ ) and **undetected** events (indexed by  $j$ )

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

- ❖ Marginalising over the unobserved data we obtain

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

$$N_{\text{ndet}}\left(\vec{\lambda}\right) \equiv \int_{\left\{\mathbf{x} < \text{threshold}\right\}} d\mathbf{x} d\vec{\theta} p\left(\mathbf{x} \mid \vec{\theta}\right) \frac{dN}{d\vec{\theta}}\left(\vec{\lambda}\right)$$

- ❖ Marginalising over the unknown number of unobserved events then gives

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

---

# Selection Effects

---

- ❖ Writing

$$\frac{dN}{d\vec{\theta}} \equiv Np(\vec{\theta} | \vec{\lambda}')$$

- ❖ and introducing a scale-invariant prior on the overall rate

$$p(N) \propto \frac{1}{N}$$

- ❖ and noting

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

- ❖ we recover the previous result.



---

# Selection Effects

---

- ❖ Typically in population analysis each event has parameters,  $\vec{\theta}$ , that we are interested in and which are drawn from the population prior with hyperparameters  $\vec{\lambda}$ .
- ❖ The “detected events” likelihood becomes

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

- ❖ The joint likelihood of data and source parameter values, 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 as before, but conditioned on source parameters

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

---

# Selection Effects

---

- ❖ The second term is the prior on source parameters *for sources above threshold* which is not the same as the population prior.

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

- ❖ Putting things together, the terms relating to selection on source parameters cancel and the joint likelihood is just

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

- ❖ The posterior on the source parameters is therefore unchanged from what you would write down if there were no selection effects. **Selection only affects inference on population parameters.**

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

---

# Hierarchical model example: fishery

---

- ❖ **Example 1:** *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  $Beta(a,b)$ .
  - We define a suitable hyperprior  $p(a,b)$  on the hyperparameters, e.g., a Gamma prior.

---

# Hierarchical model example: fishery

---

- 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  $\{p_j\}$

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

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



---

# Hierarchical model example: cosmology

---

❖ **Example 2: Cosmology with GW170817** *The first binary neutron star observed by LIGO/Virgo was also seen as a short GRB and a kilonova in the electromagnetic spectrum. This allowed the identification of a unique host galaxy, NGC 4993, and hence a determination of the source redshift. The GW observation provided a measurement of the luminosity distance and together these enable a constraint on the local expansion rate of the Universe, the Hubble constant  $H_0$ , since  $v=cz=H_0d$  locally.*

❖ We construct a hierarchical model for this measurement as follows

- The gravitational wave data,  $x_{GW}$ , depends on the source parameters, and we can marginalise over all of these except distance and inclination

$$p(x_{GW} | d, \cos \iota) = \int p(x_{GW} | d, \cos \iota, \vec{\lambda}) p(\vec{\lambda}) d\vec{\lambda}$$

- The measured recessional velocity depends on the Hubble velocity,  $H_0d$ , and the *peculiar velocity* of the host galaxy which we model as

$$p(v_r | d, v_p, H_0) = N[v_p + H_0d, \sigma_{v_r}^2](v_r)$$

---

# Hierarchical model example: cosmology

---

- The smoothed peculiar velocity field can be measured from galaxy correlations

$$p(\langle v_p \rangle | v_p) = N \left[ v_p, \sigma_{v_p}^2 \right] (\langle v_p \rangle)$$

- The combined likelihood is given by the product

$$p(x_{\text{GW}}, v_r, \langle v_p \rangle | d, \cos \iota, v_p, H_0) = \frac{1}{\mathcal{N}_s(H_0)} p(x_{\text{GW}} | d, \cos \iota) p(v_r | d, v_p, H_0) p(\langle v_p \rangle | v_p)$$

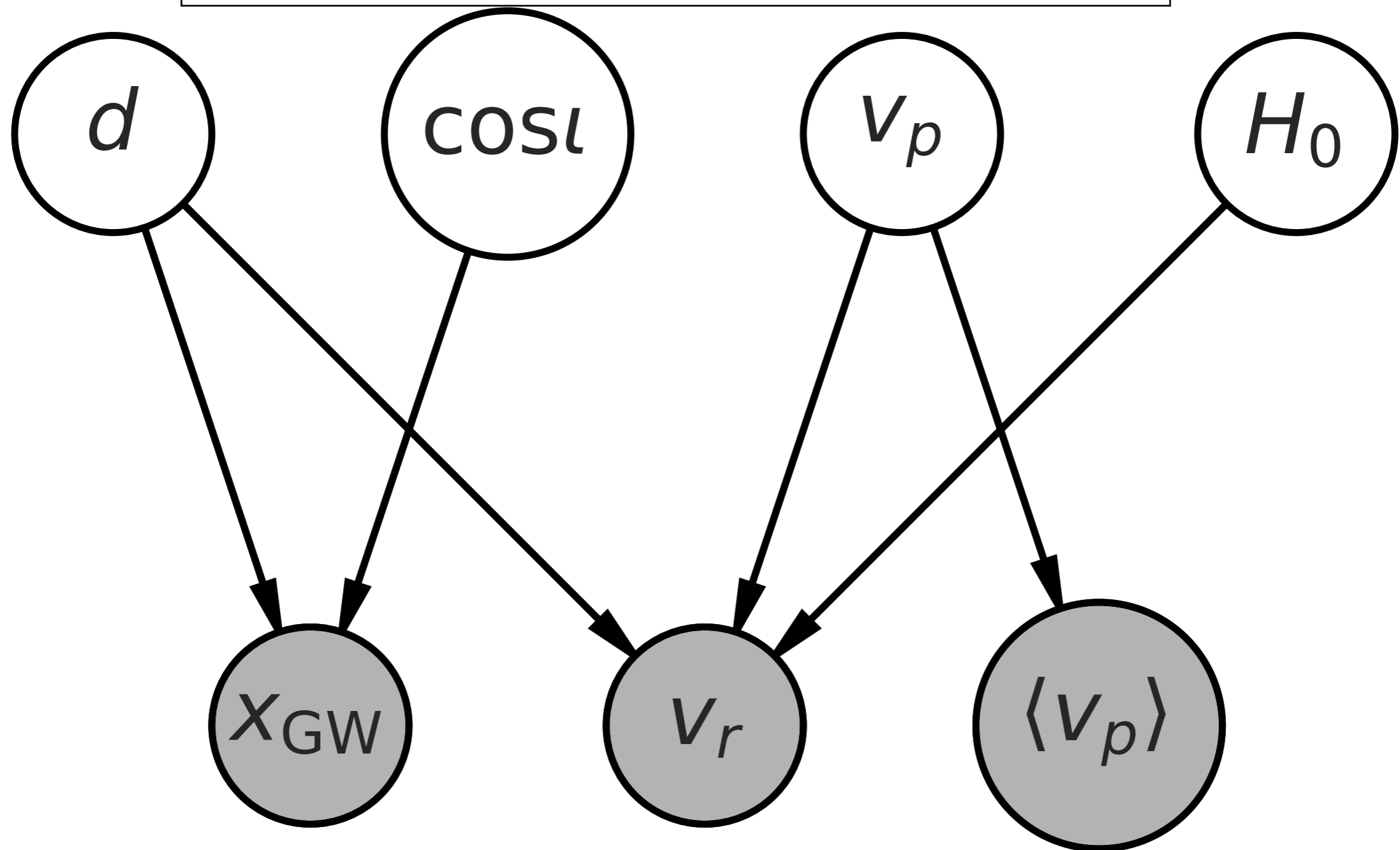
- The selection effects are encoded in

$$\mathcal{N}_s(H_0) = \int_{\mathbf{x}_{\text{GW}} > \text{threshold}} d\vec{\lambda} dd dv_p d\cos \iota d\mathbf{x}_{\text{GW}} dv_r d\langle v_p \rangle \left[ p(\mathbf{x}_{\text{GW}} | d, \cos \iota, \vec{\lambda}) p(v_r | d, v_p, H_0) \right. \\ \left. \times p(\langle v_p \rangle | v_p) p(\vec{\lambda}) p(d) p(v_p) p(\cos \iota) \right]$$

- For the GW170817 measurement the selection effect was independent of  $H_0$ .
- Finally we specify our priors  $p(d, \cos \iota, v_p, H_0) = p(d)p(\cos \iota)p(v_p)p(H_0)$

# Hierarchical model example: cosmology

Graphical model for GW170817 cosmological analysis



---

# Hierarchical model example: cosmology

---

- ❖ The posterior is

$$p(H_0, d, \cos \iota, v_p \mid x_{\text{GW}}, v_r, \langle v_p \rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} p(x_{\text{GW}} \mid d, \cos \iota) p(v_r \mid d, v_p, H_0) \\ \times p(\langle v_p \rangle \mid v_p) p(d) p(v_p) p(\cos \iota)$$

- ❖ Integrating over  $d$ ,  $v$  and  $\cos i$  we obtain the marginal posterior on  $H_0$

$$p(H_0 \mid x_{\text{GW}}, v_r, \langle v_p \rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} \int dd dv_p d\cos \iota p(x_{\text{GW}} \mid d, \cos \iota) p(v_r \mid d, v_p, H_0) \\ \times p(\langle v_p \rangle \mid v_p) p(d) p(v_p) p(\cos \iota)$$

- ❖ For multiple events the posterior becomes

$$p(H_0 \mid \{x_{\text{GW}}^i, v_r^i, \langle v_p \rangle^i\}) \propto \frac{p(H_0)}{\mathcal{N}_s^N(H_0)} \prod_{i=1}^N \left[ \int dd dv_p d\cos \iota p(x_{\text{GW}}^i \mid d, \cos \iota) p(v_r^i \mid d, v_p, H_0) \right. \\ \left. \times p(\langle v_p \rangle^i \mid v_p) p(d) p(v_p) p(\cos \iota) \right]$$

# Hierarchical model example: cosmology

