# Prediction and Bayesian model criticism/comparison

# Aims of Session 4

Tools for checking and comparing models and their practical use.

- Model checking: does my model represent the data well?
  - i.e. well enough to draw scientific conclusions? given "all models are wrong, some are useful".
  - In Bayesian analyses, need to check both prior and likelihood and their *consistency* with the data.
- Sensitivity analysis: do two alternative models give different answers?
- Model comparison: is one model better than another?
  - in terms of *consistency* with data
  - if models give same answer, inference is possibly *robust* to alternative model assumpions
  - comparison doesn't say that either of them are *good enough*

# Overview of Session 4

1. Prediction
   - Out-of-sample prediction for prediction, prediction as imputing missing data
2. Model checking: checking the likelihood
   - In-sample prediction for model checking
   - Bayesian p-values
   - Out-of-sample prediction for model checking (cross-validation)
   - Bayesian residuals for model assessment
3. Practical 4a
4. Model checking: checking the likelihood
   - Deviance summaries for model assessment
5. Checking the prior
   - Sensitivity analyses to the prior
   - Prior-data conflict
6. Model comparison
   - Penalised deviances for model comparison
7. Practical 4b

# Prior- and posterior-predictive distributions: reminder

If we have not yet observed data $y$, and we draw samples $\pi \sim Beta(a, b)$ from the prior and "replicates" $y_{rep} \sim Binomial(n, \pi)$ for each sample of $\pi$, then $y_{rep} \sim p_y(y_{rep})$ are drawn from the prior-predictive distribution (forward sampling from the prior).

In JAGS/rjags:

```
model {
  # Fixed denominator
  n <- 1000

  # Prior
  pi ~ dbeta(9.2, 13.8)

  for(i in 1:M) {

    # Prior predictive
    yrep[i] ~ dbinom(pi, n)
  }
}
```

# Prior- and posterior-predictive distributions: reminder

If we have not yet observed data $y$, and we draw samples $\pi \sim Beta(a, b)$ from the prior and "replicates" $y_{rep} \sim Binomial(n, \pi)$ for each sample of $\pi$, then $y_{rep} \sim p_y(y_{rep})$ are drawn from the prior-predictive distribution (forward sampling from the prior).

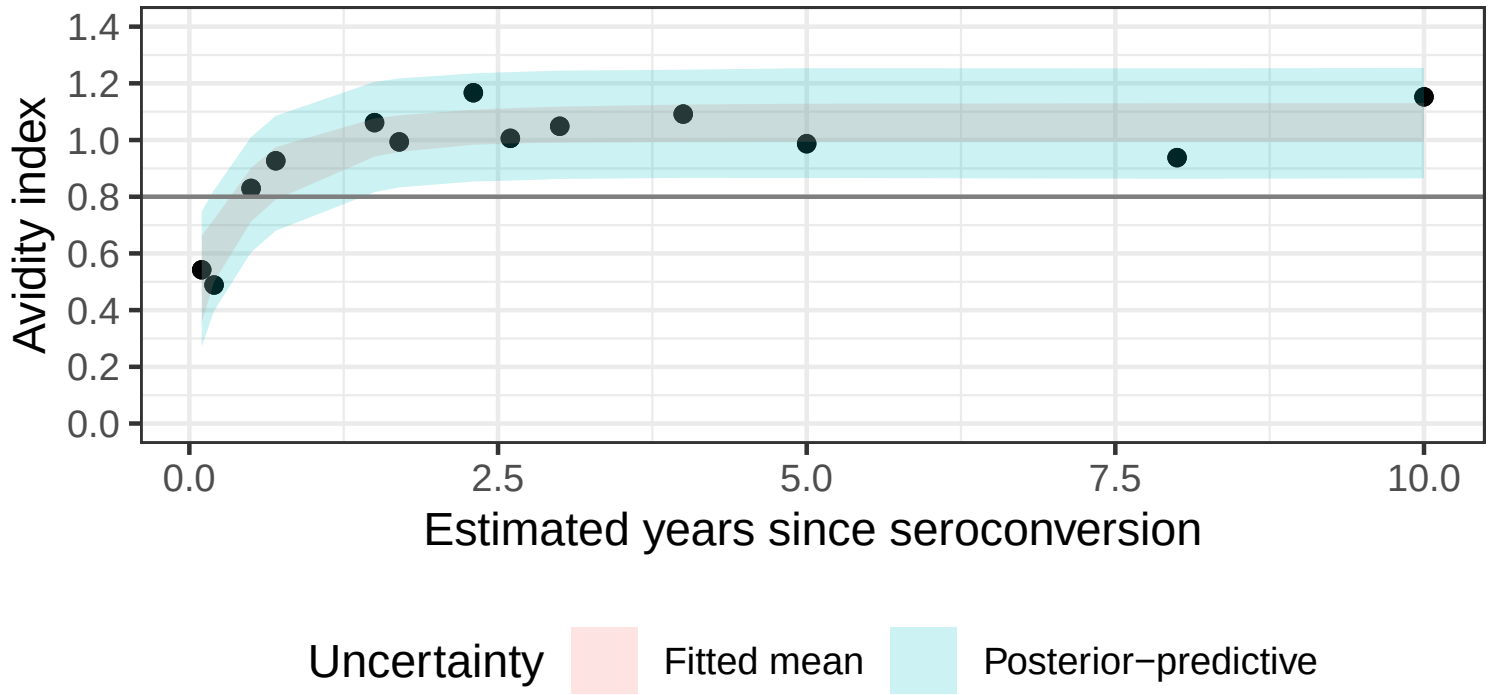
If we have observed $y$ and have obtained the posterior distribution $p_{\pi|y}(\pi|y) \propto p_{y|\pi}(y|\pi)p_\pi(\pi)$, then we draw replicates $y_{rep} \sim p_y(y_{rep})$, these replicates are drawn from the posterior-predictive distribution.

In JAGS/rjags:

```
model {
  # Fixed denominator
  n <- 1000

  # Prior
  pi ~ dbeta(9.2, 13.8)

  for(i in 1:M) {
    # Likelihood
    y[i] ~ dbinom(pi, n)

    # Posterior predictive
    yrep[i] ~ dbinom(pi, n)
  }
}
```

# HIV avidity index: prediction

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2) \text{ with } \mu_i = \mu_\infty - \mu_{diff}\gamma^{t_i}$$




Will see in next few slides how to use the predictive distribution for different purposes: prediction, imputation of missing data, model checking.

# Out-of-sample prediction for prediction I

Suppose we want to project beyond current observations, eg for the HIV avidity index, at times 0.0833 (1 month after seroconversion) and 11 years.

Recall the non-linear regression model we fit is

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2), \ i = 1, \ldots, N \text{ with } \mu_i = \mu_\infty - \mu_{diff}\gamma^{t_i}$$

where

- $\mu_\infty$ is the final size (asymptote)
- $\mu_{diff}$ is amount grown in lifetime (as time since seroconversion $t \to \infty$)
- $\gamma$ is the proportion of final size still to grow at time $t = 1$

# Out-of-sample prediction for prediction II

For implementation in JAGS/rjags, we could explicitly code predictions:

```
for (i in 1:N) {
  y[i] ~ dnorm(mu[i], prec)
  mu[i] <- mu_inf - mu_diff * gamma^t[i]
}
mu1m  <-   mu_inf - mu_diff * gamma^(1/12)
mu11  <-   mu_inf - mu_diff * gamma^11
y1m    ~   dnorm( mu1m, prec )
y11    ~   dnorm( mu11, prec )
```

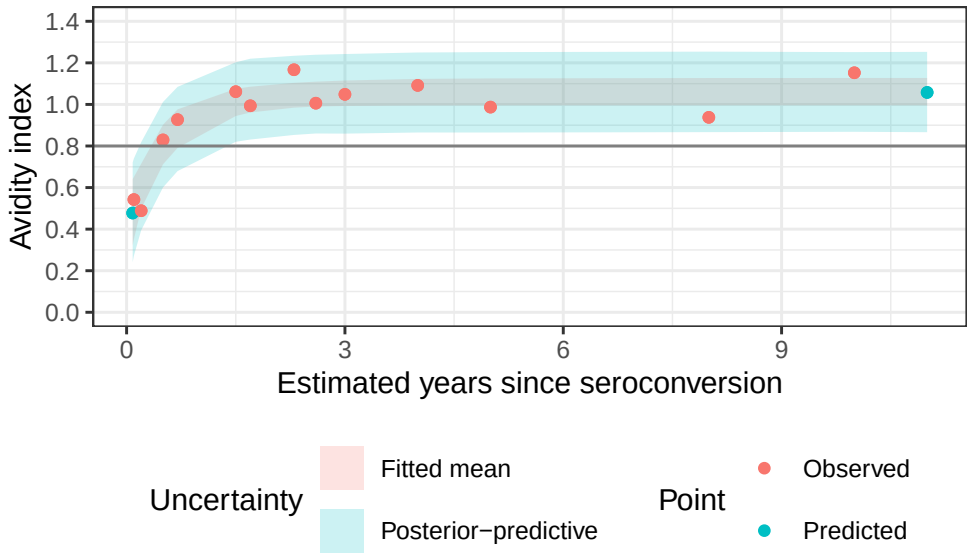
Interval around $\mu_{1m}$ will reflect uncertainty concerning fitted parameters

Interval around $y_{1m}$ will additionally reflect sampling error $\sigma = (1/prec)^{0.5}$ and uncertainty about $\sigma$.

# Prediction as imputing missing data

But it is easier to set up as missing data for the time points for which we want to predict the avidity index (see also Session 6 on missing data):

```
> hivai_miss
$t
 [1]  0.08333333  0.10000000  0.20000
 [4]  0.50000000  0.70000000  1.50000
 [7]  1.70000000  2.30000000  2.60000
[10]  3.00000000  4.00000000  5.00000
[11]  8.00000000 10.00000000 11.00000

$y
 [1]        NA 0.5424213 0.4887628
 [4] 0.8294032 0.9268438 1.0614872
 [7] 0.9936381 1.1669538 1.0060968
[10] 1.0486331 1.0916076 0.9871549
[13] 0.9379852 1.1524494        NA

$N
[1] 15
```



# Model-checking for non-hierarchical models

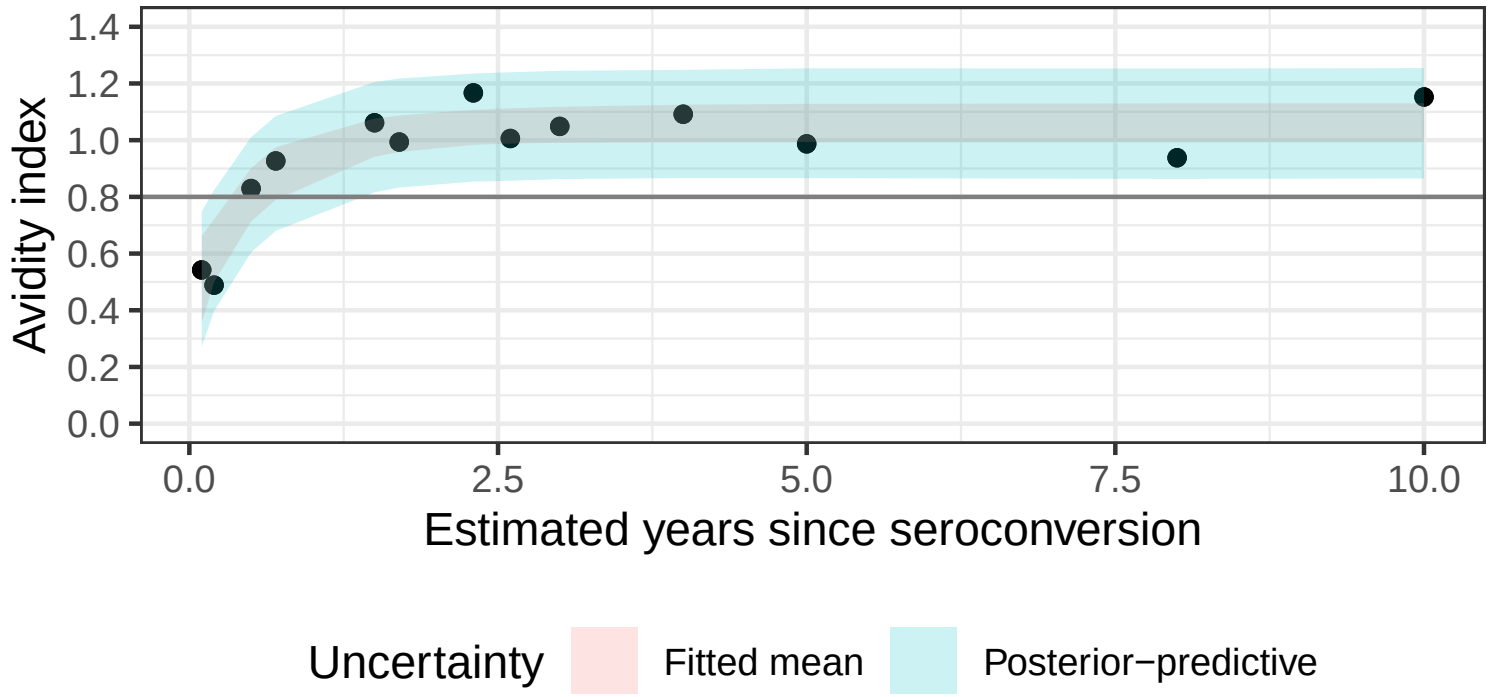
'Standard' checks that you may be familiar with from frequentist approaches, such as

- predictive checks: e.g. checking predictions against an external validation set / cross-validation
- residuals: e.g. plotting against fitted values or covariates

can all also be applied in a Bayesian framework with some key differences and additional considerations:

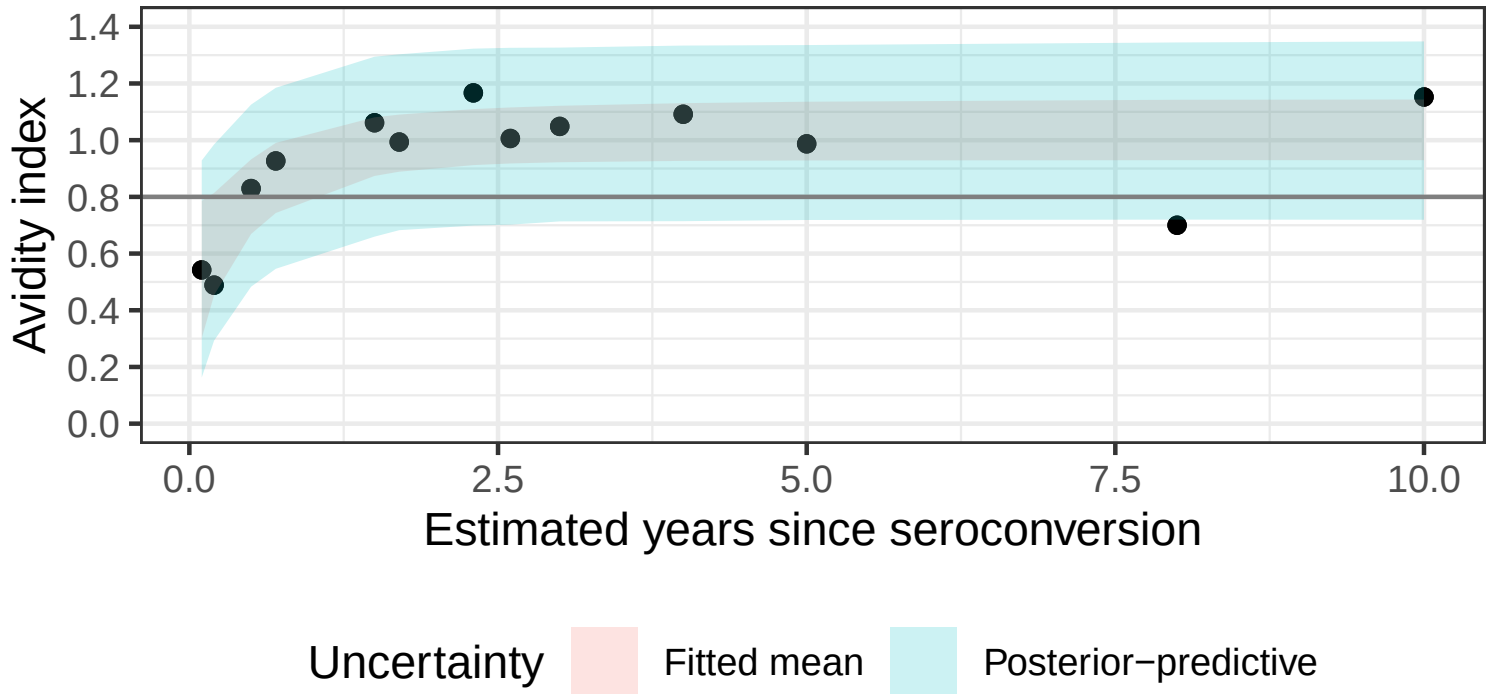
- since all parameters have distributions, residuals are also random variables with distributions
- sensitivity of estimates to the prior distribution assumed should be checked
- any potential conflict between the prior and the data should be checked (e.g. via prior-predictive checks)

# In-sample prediction for model assessment



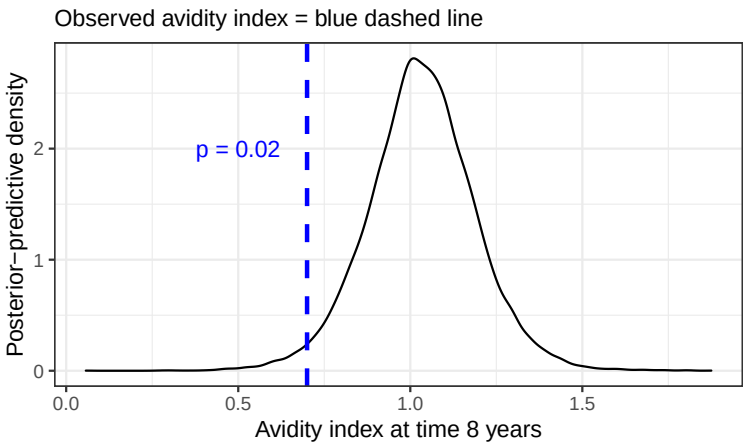

Comparing the observed data $y$ to the posterior-predictive distribution $p_Y(y)$ allows us to assess goodness-of-fit of the model to the data, e.g. via Bayesian p-values: where in the predictive distribution do the observations lie?

# In-sample prediction for model assessment




Comparing the observed data $y$ to the posterior-predictive distribution $p_Y(y)$ allows us to assess goodness-of-fit of the model to the data, e.g. via Bayesian p-values: where in the predictive distribution do the observations lie?

# Bayesian p-values I




A Bayesian p-value is the probability that a prediction (replicate) is more extreme than the observed value:

$$p = Pr\{Y_{rep} < y\}$$

If we have a sample $\boldsymbol{y_{rep}} = \{y^i_{rep}, i = 1, \ldots, M\}$ from the posterior-predictive distribution, can calculate $p$ in R simply by $\texttt{mean}(\boldsymbol{y_{rep}} < y)$ (see practical).

Under the null hypothesis $H_0$ : that the model adequately represents the data (i.e. no conflict between the prediction and the observation $y$), and *if* the likelihood dominates the prior (so that the posterior is approximately normal), then

$$p \sim \text{Unif}(0, 1)$$

# Bayesian p-values II

In the case where the observation $y$ has been used in the inference, this measure of extremity of the prediction relative to the observation is conservative, i.e. the p-values $p$ are *not* uniform, tending instead to be concentrated nearer 0.5 (Gelman, *Electron. J. Statist.*, 2013).

This conservatism arises because the observation $y$ is used twice:

- to obtain the posterior and therefore the posterior-predictive distribution;
- and in the calculation of the p-value.

# Bayesian p-values III

However, the conservatism also means that if we obtain a very small or very large p-value, then they likely *do* indicate lack of fit to the data.

All the above applies also to any function $T(y)$ of the data (e.g. the mean), known as a discrepancy statistic, used to obtain a Bayesian p-value, as a posterior-predictive check:

$$p = Pr\{T(Y_{rep}) < T(y)\}$$

# Cross-validation I

To avoid the conservatism of posterior-predictive checks, cross-validation is used so that the data are not used twice. Ideally, we should check the model with new data

- $y_f$ - model is fit to the original data (training set)
- $y_c$ - new data (validation set) are used to assess the model fit

## Out-of-sample prediction

1. Obtain the posterior distribution $p(\theta \mid y_f)$
2. Obtain the posterior-predictive distribution for the new data

$$p\left(y_c^{\text{rep}} \mid y_f\right) = \int p\left(y_c^{\text{rep}} \mid y_f, \theta\right) p\left(\theta \mid y_f\right) \, d\theta$$
$$= \int p\left(y_c^{\text{rep}} \mid \theta\right) p\left(\theta \mid y_f\right) \, d\theta$$

3. Compare $y_c^{\text{rep}}$ and $y_c$ using Bayesian p-value $p = Pr\{T(Y_c^{rep}) < T(y_c)\}$

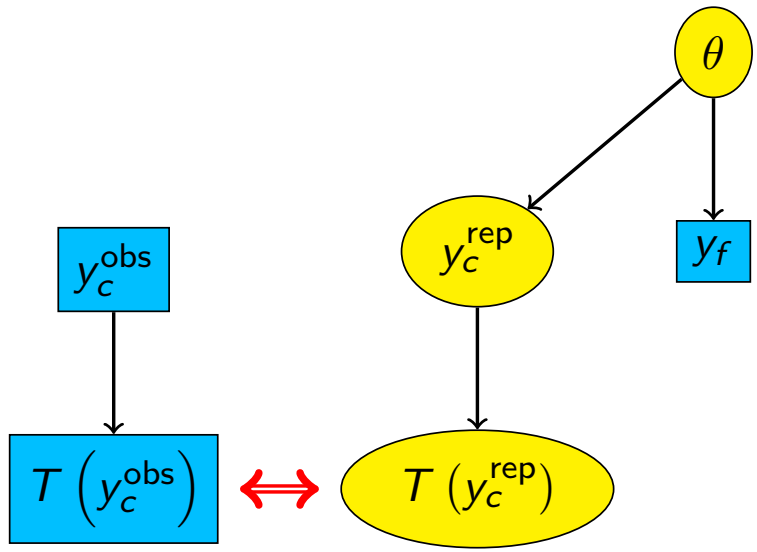
# Cross-validation II

But often, no new data available, so we split the available original data into $y_f$ and $y_c$.

## Splitting options

Leave-one-out cross-validation Set $y_c = y_i$, the $i$'th single observation, and $y_f = y_{\text{I}}$, the rest. Repeat for each observation $i$, leaving one out at a time.

$k$-fold cross-validation Set $y_f$ to be $k$% of the data and $y_c$ to be the remaining $(100 - k)$%. Repeat for each 'fold', leaving $(100 - k)$% out at a time.




As before, comparison of the data and the predictive distribution is carried out via a Bayesian p-value. See Session 5 for an example of cross-validation in hierarchical models.

# Comments on cross-validation

- computationally intensive to implement using MCMC
- does not necessarily target specific model assumptions
- multiple testing issue - use e.g. False Discovery Rates

Alternatives to cross-validation:

- importance resampling (not covered in this course)
- approximate $p(Y_i^{pred}|y_{\setminus i})$ by $p(Y_i^{pred}|y)$ (conservative)
  - just replicate data ('posterior predictive P-values')
  - replicate parameters *and* data ('mixed predictive P-values', see also Session 5 Hierarchical models)

# Bayesian residuals I

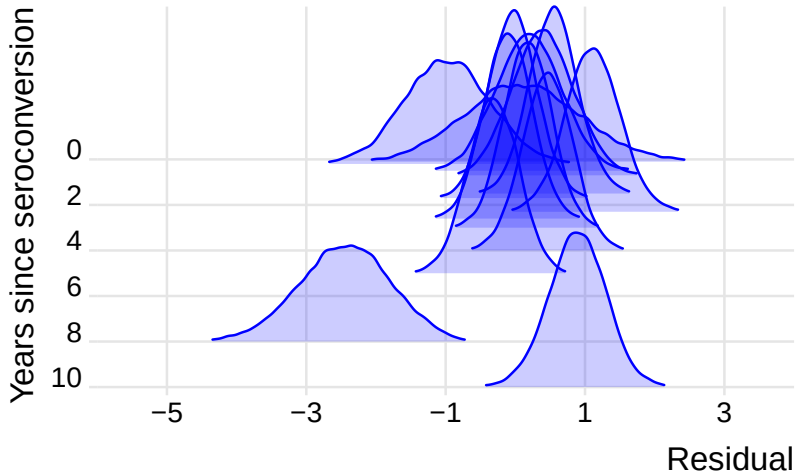
Standardised Pearson residuals are defined as:

$$\rho = (y - \mu)/\sigma$$

where $\mu = \mathbb{E}(Y), \sigma^2 = \text{Var}(Y)$.

Since $\mu$ and $\sigma$ are random variables with distributions, so are the residuals $\rho$.

If we have assumed a Normal likelihood $Y \sim N(\mu, \sigma^2)$, then $\rho \sim N(0, 1)$, so we might broadly expect the residuals to lie between -2 and 2.



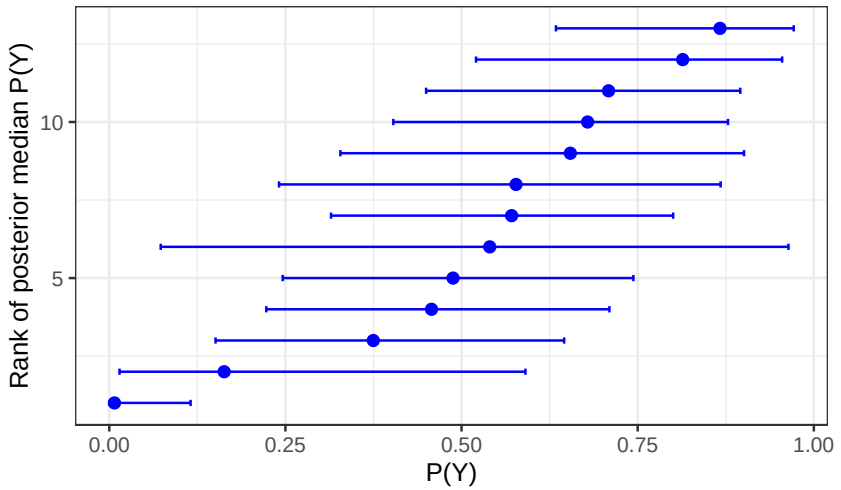
# Bayesian residuals II

In the frequentist framework, with fixed $\mu$ and $\sigma$, the associated p-values

$$P(Y) = \Phi[(Y - \theta)/\sigma]$$

are uniformly distributed $P(Y) \sim$Unif$(0, 1)$.

- Might therefore look for linearity in a plot of the *ranked* p-values (similar to Q-Q plots).
- *But* $\mu, \sigma$ are random variables in the Bayesian framework, so an informal exploratory diagnostic only.



# Practical 4a

- Will use the HIV avidity index example to explore prediction
  - for prediction
  - for model assessment
- as well as Bayesian residuals

aiming to:

- understand the role of prediction and it's use in both in- and out-of-sample prediction for the different purposes above;
- practice manipulating posterior samples after having obtained them to derive various quantities of interest, such as residuals and posterior-predictive p-values, to aid in model checking;
- practice plotting credible and prediction intervals and other posterior summaries to aid visually in model checking.

# Deviance summaries I

The deviance is defined as twice the minus log-likelihood:

$$D(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}$$

and is a general measure of how well the model fits the data used for estimating the model parameters (i.e. in-sample fit, rather than out-of-sample predictive ability).

## Example: Binomial

If we assume a Binomial sampling distribution $\boldsymbol{y} \sim \text{Binomial}(n, \pi)$ for a set of $m$ observations $\boldsymbol{y} = y_1, \ldots, y_m$, then the deviance is

$$D(\pi) = -2\sum_{i=1}^{m}\left[y_i \log(\pi) + (n - y_i)\log(1-\pi) + \log\binom{n}{y_i}\right]$$

# Deviance summaries II

As with the residuals, since the deviance is a function of the parameters, which are random variables, the deviance also has a distribution.

In `JAGS/rjags`, the deviance (along with other model assessment and comparison measures) can be obtained (for a JAGS model `model` with at least 2 chains) by:

```
load.module("dic")
model_dic <- dic.samples(model, type = "pD", n.iter = nIter)
```

The deviance is used as a component of a model comparison statistic known as the Deviance Information Criterion (see model comparison slides 35 onwards).

But for assessing absolute goodness of fit, particularly for the exponential family, a standardised deviance is more useful (not available in `dic.samples`, has to be coded either in `JAGS/rjags` or `R`).

# Standardised & saturated deviances I

## Definition: standardised deviance

$$D_S(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\} + 2\log p(\boldsymbol{y})$$

The standardised deviance is the deviance standardised by a function $p(\boldsymbol{y})$ of the data alone.

## Definition: saturated deviance

In generalised linear models (exponential family), the saturated deviance is the deviance standardised by the likelihood of the saturated model, i.e.

$$D_S(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\} + 2\log p\left(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_S(\boldsymbol{y})\right)$$

where $\hat{\boldsymbol{\theta}}_S(\boldsymbol{y})$ is an appropriate saturated estimate of $\boldsymbol{\theta}$, such as the maximum likelihood estimate, e.g. if $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\theta}$, then we use $\hat{\boldsymbol{\theta}}_S(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Note that $D_S(\boldsymbol{\theta})$ is the log-likelihood ratio of the model compared to the saturated model.

# Standardised & saturated deviances II

## Examples

i.i.d Binomial samples $\boldsymbol{y} = y_i, i = 1, \ldots, m \sim \text{Binomial}(n, \pi)$:

$$D_S(\pi) = -2 \sum_{i=1}^{m} \left[ y_i \log \frac{\pi}{y_i/n} + (n - y_i) \log \frac{(1-\pi)}{(1-(y_i/n))} \right]$$

i.i.d Poisson counts $\boldsymbol{y} = y_i, i = 1, \ldots, m \sim \text{Poisson}(\mu)$:

$$D_S(\mu) = -2 \left\{ \left( \sum_{i=1}^{m} y_i - m\mu \right) - \sum_{i=1}^{m} y_i \log \left[ \frac{\sum_{i=1}^{m} y_i}{m\mu} \right] \right\}$$

i.i.d Normal observations $\boldsymbol{y} = y_i, i = 1, \ldots, m \sim \text{Normal}(\mu, \sigma^2)$:

$$D_S(\mu, \sigma) = m \log \left( \frac{\sigma^2}{\hat{\sigma}^2} \right) + \frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - \mu)^2 - \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{m} (y_i - \hat{\mu})^2$$

where $\hat{\mu} = \frac{1}{m} \sum_i y_i$ and $\hat{\sigma}^2 = \frac{1}{m} \sum_i (y_i - \hat{\mu})^2$ are the MLEs.

# Standardised & saturated deviances III

- These expressions (and similarly for any other sampling distribution in the exponential family) can be seen as a sum of squared deviance residuals,

$$D_S(\boldsymbol{\theta}) = \sum_{i=1}^{m} D_{Si}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \delta_i^2$$

  each of which is, again, a random variable with a distribution.

- The deviance residuals $\delta_i = \pm\sqrt{D_{Si}(\boldsymbol{\theta})}$ (with sign given by the sign of $y_i - \mathbb{E}(Y_i)$) are defined analogously to the frequentist version of McCullagh & Nelder (1989).

Dempster (1974) suggested plotting the posterior distribution of $D_S(\boldsymbol{\theta})$, others have suggested considering the posterior mean deviance

$$\overline{D} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(D_S(\boldsymbol{\theta})\right)$$

# Standardised & saturated deviances IV

For the exponential family, when a model is true and under standard regularity conditions (Spiegelhalter et al, 2002):

$$\mathbb{E}_Y(\overline{D}) = \mathbb{E}_Y \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \left( D_S(\boldsymbol{\theta}) \right) \approx m$$

where $m$ is the dimension of $\boldsymbol{y}$.

$\Rightarrow$ informally assess model fit by comparison of $\overline{D}$ to the number of data points $m$ or of $\overline{D}_i$ to 1.

The contributions of each observation to posterior expected deviance are an alternative to posterior-predictive p-values for finding lack of fit, that are not conservative, and that you might want to calculate anyway as part of calculating the DIC (see model comparison slides 35 onwards).

# Beta-binomial drug example I

Recall we assumed a prior $\pi \sim \text{Beta}(9.2, 13.8)$ such that the prior mean was 0.4, prior sd 0.1 and 95% prior interval (0.2,0.6).

Suppose in a series of $m$ trials of the drug in different groups, each of size $N = 20$, we observe the following numbers of positive responses:

## Example

```
> ## Fix probability and denominator
> pi <- 0.6
> N <- 20
> m <- 10
>
> ## Simulate m data points
> (y <- rbinom(n = m, size = N, prob = pi))
 [1] 11 11 11 10 11 13 13 14  8 14
```

# Beta-binomial drug example II

We can fit the model to these simulated data in `JAGS/rjags`, and also calculate the saturated deviance at each MCMC iteration:

## Model

```
model {
  # Prior
  pi ~ dbeta(9.2, 13.8)

  for(i in 1:m) {
    # Likelihood
    y[i] ~ dbinom(pi, N)

    # Binomial saturated deviance
    dev.pi[i] <- -2 * ((y[i] * (log(pi) - log(y[i]/N))) +
        ((N-y[i]) * (log(1-pi) - log(1 - (y[i]/N)))))
  }
}
```

# Beta-binomial drug example III

or alternatively calculate the deviance in R after running the model, using the posterior samples of $\pi$ (see practical).

Taking the mean over the posterior samples, we obtain the following posterior mean deviance contributions for each observation and posterior mean estimate of yhat $= \pi N$. The posterior mean, median and 95% credible interval for $\pi$ is 0.561, 0.561 (0.496-0.626).

**Posterior mean deviance**

```
# A tibble: 10 x 3
       y  yhat   dev
   <dbl> <dbl> <dbl>
 1    11  11.2 0.102
 2    11  11.2 0.102
 3    11  11.2 0.102
 4    10  11.2 0.395
 5    11  11.2 0.102
 6    13  11.2 0.742
 7    13  11.2 0.742
 8    14  11.2 1.71
 9     8  11.2 2.19
10    14  11.2 1.71
```

# More on saturated deviance

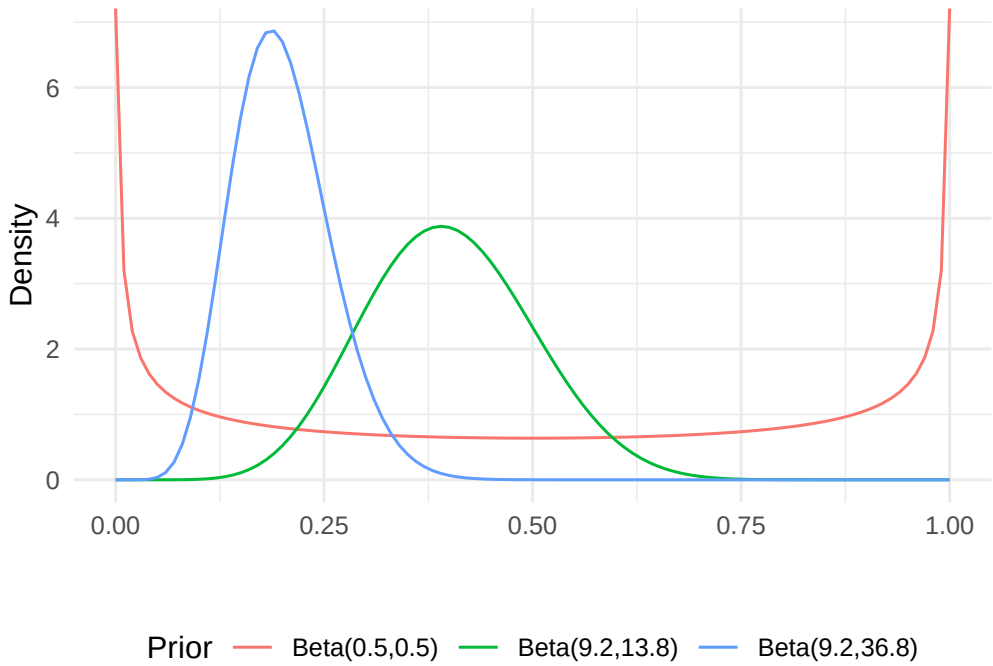
- Note that since the distribution of posterior deviance is not well understood, we can't say how far away from $m$ or 1 $\overline{D}$ or $\overline{D_i}$ need to be to count as lack of fit $\Rightarrow$ exploratory diagnostic only.
- Invariant to parameterisation of $\boldsymbol{\theta}$
- Robust, generally converges well
- But more complex models will fit the data better and so will have smaller $\overline{D}$
- Need to have some measure of 'model complexity' to trade off against $\overline{D}$ (see model comparison slides 35 onwards)

# Prior sensitivity I

Returning to the Beta-binomial drug example, suppose we have prior information from two different studies:

- one suggests around 40% (20 – 60%) of individuals respond positively to the drug (previously seen prior);
- the second suggests only 20% (10 – 33%) respond.

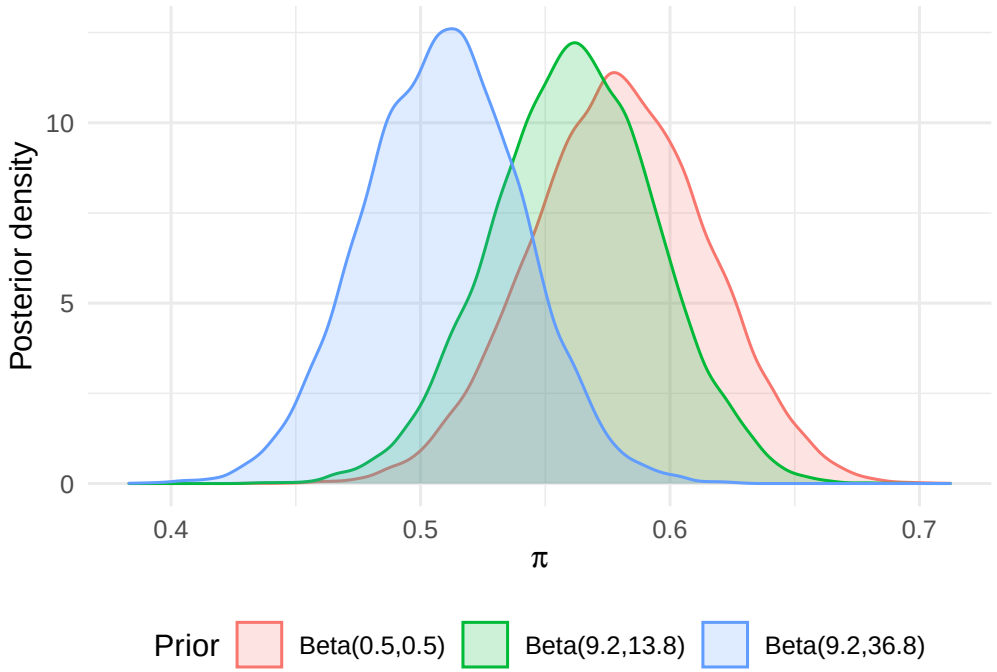
We will consider how different the posterior estimates are with priors corresponding to each, as well as a less informative prior, the three priors are plotted here:



# Prior sensitivity II

Highlights importance of assessing sensitivity to prior, as well as potential for prior-data conflict: the posterior corresponding to the vague Beta(0.5,0.5) prior (red) we expect reflects more the likelihood, suggesting the blue Beta(9.2,36.8) prior may be inconsistent with the data.

Each prior results in a somewhat different posterior distribution for $\pi$.



# Prior-data conflict checking via prior-predictive checks I

We can carry out prior-predictive checks to check for conflict between our assumed prior and the data.

## Posterior-predictive

```
model {
  # Prior
  pi ~ dbeta(9.2, 36.8)

  for(i in 1:m) {
    # Likelihood
    y[i] ~ dbinom(pi, N)

    # Posterior predictive
    yrep[i] ~ dbinom(pi, N)
  }
}
```

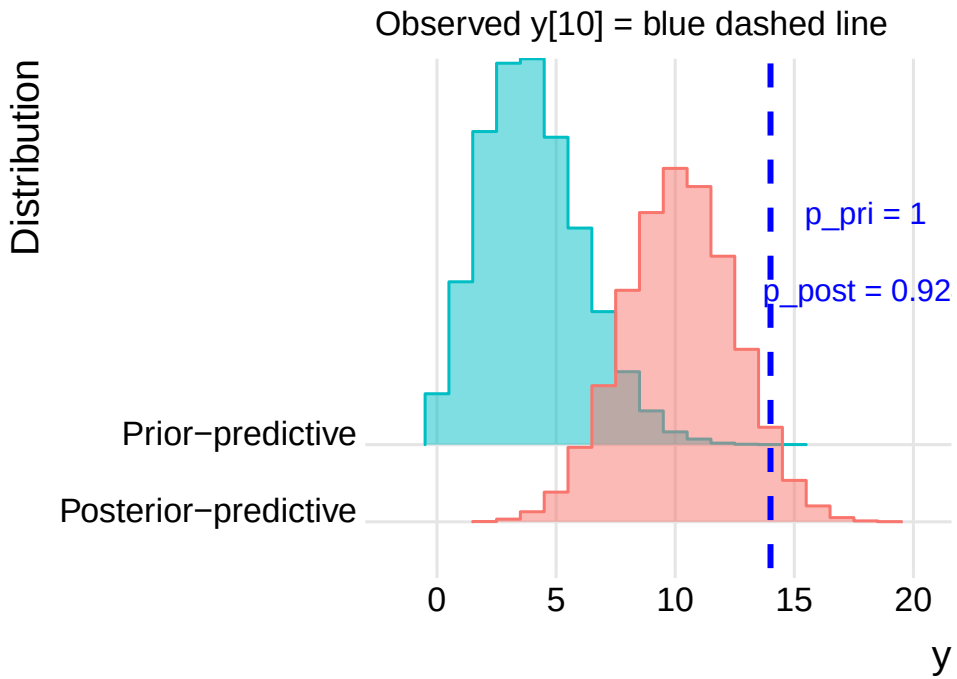
## Prior-predictive

```
model {
  # Prior
  pi ~ dbeta(9.2, 36.8)

  for(i in 1:m) {
    # Prior predictive
    yrep[i] ~ dbinom(pi, N)
  }
}
```

# Prior-data conflict checking via prior-predictive checks II

Or alternatively, in R:

```
# Prior
pi_pri <- rbeta(n = 20000,
                shape1 = 9.2,
                shape2 = 36.8)

# Prior-predictive
yrep_pri <- sapply(pi_pri, function(x) {
  rbinom(n = 1, size = N, prob = x)
  })
```



# Model comparison I

Recall that we want to assess how well a model predicts outside the data.

- More complex models might fit the observed data better (likelihood), but could be worse at generalising to other data.
- So various model comparison measures have been derived, which all approximate different kinds of cross-validation.
- These measures combine a measure of model fit, penalised by a measure of model complexity.

# Model comparison II

## Model complexity

Spiegelhalter et al (2002) proposed to measure complexity using the effective number of parameters:

$$p_D = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(D) - D(\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})) = \overline{D} - D(\overline{\boldsymbol{\theta}})$$

where $D(\overline{\boldsymbol{\theta}})$ is the plug-in deviance, i.e. the deviance function evaluated at the posterior mean (or median) of the parameters.

- Depending on the amount of prior information, $p_D$ is often *smaller* than the actual number of parameters, since informative priors constrain the inference.
- Plummer (2008) proposes an alternative measure of complexity, the optimism $p_{\text{opt}}$, which penalises model complexity more severely than $p_D$. Both are available in `JAGS/rjags`, using the unstandardised deviance.

# Model comparison III

## Deviance Information Criterion

To compare models, Spiegelhalter et al (2002) proposed the DIC, analogously to Akaike's Information Criterion, as:

$$\begin{aligned} \text{DIC} &= \text{'goodness of fit'} + \text{'complexity'} \\ &= D(\overline{\boldsymbol{\theta}}) + 2p_D \\ &= \overline{D} + p_D \end{aligned}$$

Plummer (2008) instead propose the penalised expected deviance $\overline{D} + p_{\text{opt}}$.

Models with smaller DIC are better supported by the data. Very roughly:

- differences in DIC $> 10$ are substantial
- differences in DIC $< 5$ may be negligible

# Comments on model comparison

- Both the Spiegelhalter et al (2002) and Plummer (2008) versions are available in `JAGS/rjags` using `dic.samples()` (`type = "pD"` or `type = "popt"`), again using the unstandardised deviance.
- $p_D$ is not invariant to reparameterisation, and can sometimes be negative (not interpretable as an effective number of parameters).
- Both versions of penalised deviances are approximations to different cross-validatory loss functions
- See also the Widely Applicable Information Criterion (WAIC), LOO and the `R loo` package (Vehtari, Gelman and Gabry, Stat. Comp. 2017), which give a more direct approximation to leave-one-out cross-validation, easily computed in a wide range of situations (not covered in this course).

# HIV AI example: model comparison I

Returning to the HIV AI example, let's consider fitting a $t$-distribution to the dataset with one artificial outlier, instead of a Normal distribution.

## Student's $t$ distribution

The probability density function is

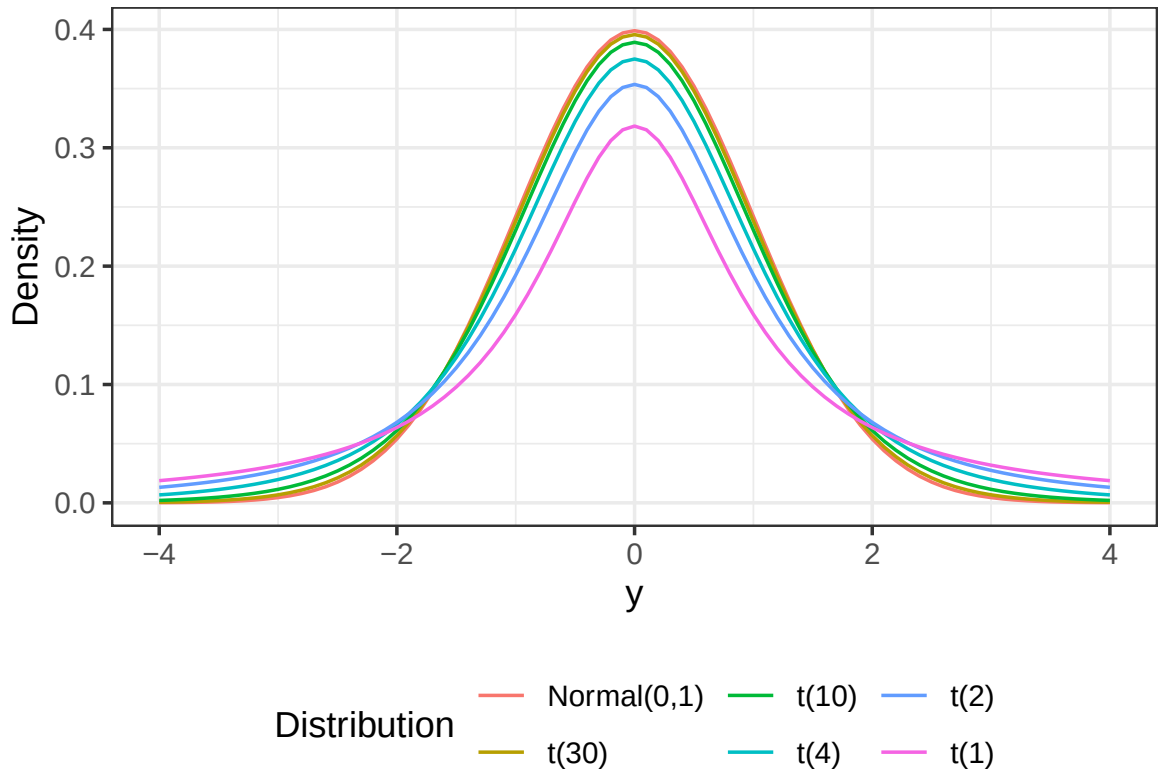
$$p(y; \mu, \tau, k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{\tau}{k\pi}\right)^{\frac{1}{2}} \left\{1 + \frac{\tau(y-\mu)^2}{k}\right\}^{-\frac{(k+1)}{2}}$$

where $\mu$ is the location/mean; $\tau$ is the precision; $k$ is the number of degrees of freedom; and $\Gamma$ is the gamma function.

- We can specify a $t$ sampling distribution in `JAGS/rjags` as `y ~ dt(mu, tau, k)`.

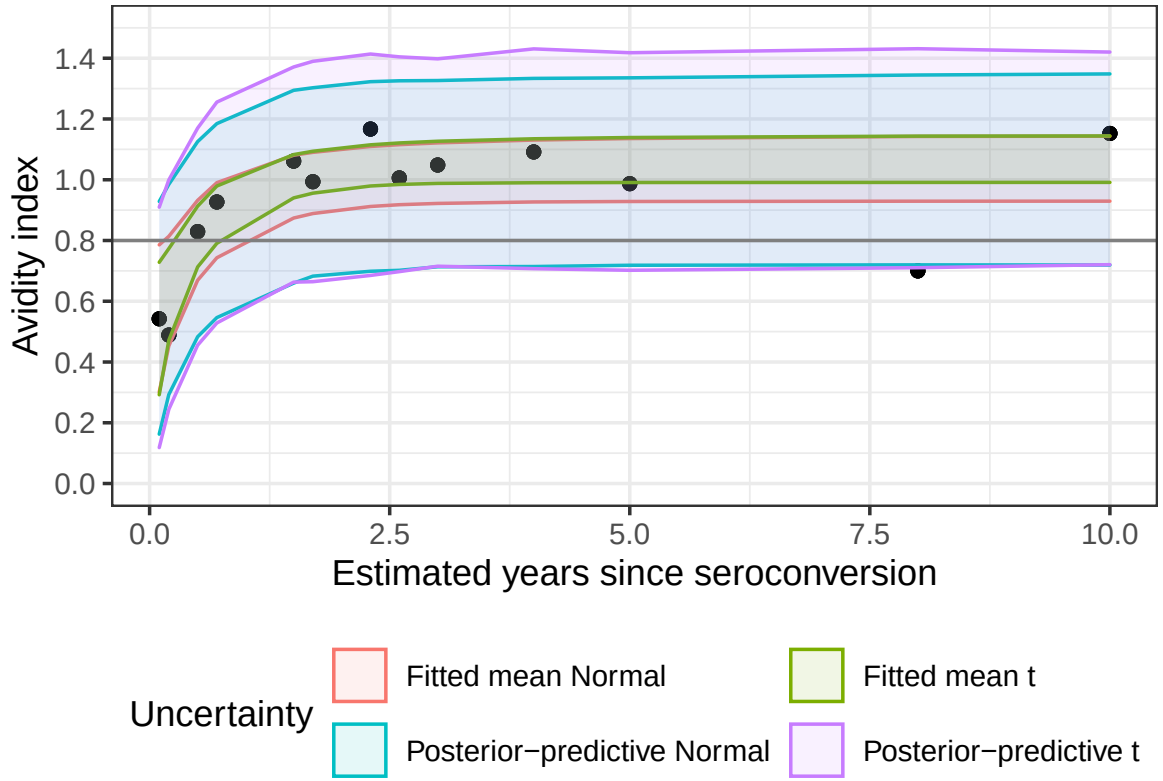
# HIV AI example: model comparison II

- The $t$-distribution has heavier tails than the Normal, hence more likely to accommodate outliers.
- As $k$ increases, the $t$-distribution approaches a $N(0, 1)$ distribution ($k = 30$ is a good approximation).



# HIV AI example: model comparison III

In theory, it is possible to estimate the df $k$, but in practice, challenging unless you have a huge amount of data, with enough in the tails. So we fix e.g. $k = 2$ in the HIV AI example.



# HIV AI example: model comparison IV

Penalised deviances (DIC and Penalised Expected Deviance - PED) obtained from the `JAGS/rjags dic.samples()` function:

| Distribution | Type | Posterior mean deviance | Penalty | Penalised deviance |
|---|---|---|---|---|
| Normal | DIC | -14.51 | 4.62 | -9.90 |
| $t$ | DIC | -17.79 | 4.07 | -13.73 |
| Normal | PED | -14.36 | 12.49 | -1.87 |
| $t$ | PED | -18.14 | 10.63 | -7.51 |

# Practical 4b

- Will use the Beta-Binomial drug example to explore deviance residuals, prior sensitivity and prior-data conflict
- Will use the HIV AI example to explore model comparison

aiming to:

- understand the role of prediction and it's use in both in- and out-of-sample prediction for prediction, model checking, model comparison;
- practice manipulating posterior samples after having obtained them to derive various quantities of interest, such as deviance summaries and prior-predictive p-values, to aid in model checking and comparison;
- practice plotting credible and prediction intervals and other posterior summaries to aid visually in model checking and comparison;
- practice thinking about different aspects of model fit, sensitivity analysis and model comparison.

# Summary

- Prediction important for multiple reasons
- Model checking should become a default part of your model-building process
  - cycle of model development $\Rightarrow$ model checking $\Rightarrow$ model comparison