Introduction to Generalised Linear Models for Ecologists

Dr Niamh Mimnagh

niamh@prstats.org

https://github.com/niamhmimnagh/ GLME01---Introduction-to-Generalised-Linear-Models-for-Ecologists

PR STATS

Bayesian Statistics

- Bayesian statistics is a way to learn from data by updating what you already know into your current beliefs about unknowns.
- It works by treating parameters as uncertain, and talks in probabilities about effects.
- We begin with a prior (what we believe before seeing any data), and we combine it with our likelihood (modelling the data) and we end up with a posterior (what we believe after seeing the data).
- We use Bayesian statistics because it gives us decision-ready probabilities and stable estimates when data are small or tricky

Issues Bayesian Methods Address

- Small samples or rare events: estimates can be jumpy; standard errors blow up; conclusions feel fragile.
- Separation in logistic regression: when a predictor almost perfectly splits 0 vs 1, maximum likelihood struggles or fails; weakly informative priors stabilise estimates.
- Extreme probabilities and skewed counts: near - 0/1 probabilities and long-tailed or zero-heavy counts benefit from regularisation.
- Interpretability: stakeholders ask: “What’s the probability the effect is beneficial?” rather than “Is $p < 0.05$?”
- Bayesian contribution: encode reasonable starting beliefs, update with data, and report clear probability statements about effects and predictions.

Two Ways to Think About Uncertainty

- Repeatability view (frequentist): imagine rerunning your study many times under identical conditions; procedures are designed to perform well across those hypothetical repetitions.
- Belief view (Bayesian): given the data in front of us and what we reasonably believed beforehand, how plausible are different values of the effect now?
- Both are coherent and useful; they answer different questions. Bayesian answers often match how people naturally talk about uncertainty.

Frequentist Inference

- Parameters are fixed but unknown; data are random. We ask how our method behaves over many hypothetical repeats.
- Confidence intervals are procedural: a 95% CI is a technique that would capture the true value 95% of the time across repeated samples (it is not a 95% probability about this one interval).
- Hypothesis testing language: “If there were no effect, would these data be unusually large or small?”
- Strengths: widely used, fast to compute, strong long-run guarantees.
- Limits: hard to express as the probability an effect is positive or beneficial in the current study.

Bayesian Inference

- Parameters are uncertain; data are what we observed.
- We combine two ingredients to form what we believe now (the posterior):
  - Prior — what we thought before we saw this dataset.
  - Likelihood — how compatible the observed data are with different parameter values.
- Interpretation payoff: “There is an 88% probability that the treatment increases the success rate.”
- Strengths: direct probability statements, regularisation via priors, natural handling of complex models.
- Responsibilities: choose and justify priors; check model fit carefully.

How Inferences Differ

- Frequentist phrasing: “The 95% confidence interval for the odds ratio is [0.9, 1.6]; the p-value is 0.12. We fail to reject the null at 5%.”
- Bayesian phrasing: “Given the data and our prior, there is a 78% probability that the odds ratio exceeds 1, and a 62% probability that it exceeds 1.2.”
- Why this matters: decision-makers can weigh probabilities directly against costs and benefits - e.g., “Is a 78% chance of benefit enough to deploy the intervention?”

GLMs

- A GLM links predictors to outcomes that are not continuous and unbounded, using a link function appropriate for the outcome type.
- Binary outcomes: logistic regression maps a linear predictor into a probability in [0, 1], avoiding impossible predictions like −0.3 or 1.4.
- Counts: Poisson (and variants) model non-negative integers and allow variance to grow with the mean.
- The model form is the same for frequentist and Bayesian GLMs; the difference is how we quantify and interpret uncertainty.

Where Bayesian Thinking Comes In

- We use the same formula, but with a new layer.
- We place prior distributions (‘priors’) on coefficients (and other parameters), expressing what is reasonable before seeing the data.
- The data is then allowed to refine our beliefs. Thelikelihood pulls priors toward values that better explain what we observed.
- The end result is a posterior distribution for each coefficient and for predictions, not just a single number with a standard error.

The Likelihood: Listening to the Data

- The likelihood is a way of asking: "If the model's parameters had certain values, how likely is it that we would see the data we actually observed?"
- It is the bridge between data and model. It tells us how well different parameter values "explain" the data we saw.
- We start with the model:

$$Y_i \sim Poisson(\lambda_i), \text{and } \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots$$

- The likelihood is:

1. Take all the observed data values $Y_1, Y_2, \ldots, Y_n$.
2. For each observation, ask: "If the coefficients were these values, how plausible is it that we would see this Y_i"?
3. The product of those plausibilities is the likelihood.

The Prior: A Reasonable Starting Point

- A prior distribution is a way of expressing what we believe about the data before we see it.
- Ìt is not a way to force the result you want; it is a transparent statement of what values are reasonable before seeing this dataset.
- Good sources of information for the prior are: past studies, domain knowledge, physical limits, or weakly informative defaults that gently rule out extremes.

The Prior: A Reasonable Starting Point

- An informative prior is narrow. It encodes strong, specific beliefs that meaningfully constrain a parameter to a plausible range.
- A weakly informative prior is broader. It rules out absurd values, while letting the data speak.
- A flat/non-informative prior tries to express 'no information'.

- Poisson regression coefficient β:
  - Informative prior: beta ~ Normal(0, 0.5)
  - Weak prior: beta ~ Normal(0, 1)
  - Flat prior: beta ~ Uniform(0, 10)

The Prior: A Reasonable Starting Point

Informative prior: $\beta \sim Normal(0, 0.5)$

- 95% of the prior mass for β is between –1 and +1.
- Exponentiating gives multipliers between about 0.37 and 2.7.
- Before seeing data, we believe a 1-unit increase in x would most likely change the expected count to between one-third and nearly three times as large.

Weak prior: $\beta \sim Normal(0, 1)$

- 95% of the prior mass is between –2 and +2.
- Exponentiating gives multipliers between about 0.14 and 7.4.
- A 1-unit increase in x could shrink the expected count to one-seventh its size, or boost it about sevenfold.

Flat prior: $\beta \sim Normal(0, 10)$

- 95% of the prior mass is between –20 and +20.
- Exponentiating gives multipliers between about 2×10^{-9} and 5×10^{8}.
- A 1-unit increase in x could virtually wipe out the expected count or blow it up hundreds of millions of times.

How Informative are Your Priors?

- Non-informative (very flat): sounds neutral but can misbehave on odds/ratio scales, allowing absurd values to get undue weight.
- Weakly informative: gentle regularisation toward reasonable ranges (often the best default for GLMs; prevents wild swings with thin data).
- Informative: justified when strong prior knowledge exists (meta-analyses, physics).
- Principle: pick the weakest prior that still rules out nonsense; document the choice.

Regularisation: Why Priors Stabilise Estimates

- With small n MLE coefficients can be huge or undefined, and standard errors explode.
- A prior encodes that gigantic effects are unlikely. When combined with the data (likelihood), the posterior becomes a compromise that pulls back extreme, poorly supported estimates while still moving strongly when the data are persuasive.
- Priors help most for small/imbalanced datasets, highly correlated predictors, and hierarchical settings. This improves out-of-sample performance.

Where Bayesian Thinking Comes In

$$Y_i \sim Binomial(\pi_i)$$
$$logit(\pi_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

Likelihood

- We now add priors to our model. What parameters get priors?
  - The intercept β_0 slopes $\beta_1, \ldots, \beta_p$
  - If hierarchical: group effects.
  - Any extra parameters (e.g., overdispersion ϕ, zero-inflation probability in ZIP).
- What do we *not* put priors on?
  - The observed data Y_i and predictors x_{ij}
  - π_i (it's a deterministic function of β and x).

The Posterior: Updated Belief After Seeing Data

- The Posterior is a combination of the prior and the likelihood.
- We started with our prior beliefs about the parameters, then we fitted our model, and the posterior represents our updated certainty about the parameters after we've seen the data.
- What you get from the posterior:
  - Distributions for coefficients (not just single estimates).
  - Credible intervals that say "there's a 95% probability the effect lies here."
  - Predictive distributions for new observations or scenarios.

Posterior vs. MLE

Frequentist MLE (Maximum Likelihood point estimate)

- What you get: a single best-fitting $\hat{\beta}$, uncertainty via standard errors and a p-value. If the p-value is less than 0.05, congratulations!
- Strengths: fast; no priors required; well-known theory (consistency, asymptotic efficiency); great baseline.

Bayesian posterior (distribution):

- What you get: a large set of samples of plausible β that match the data best; summarise with posterior mean/median/mode, credible intervals, and posterior predictive for new data.
- Strengths: direct probabilities about effects (e.g., $P(\beta_j > 0 \mid y)$) can regularise and encode knowledge via priors; natural with small/awkward data and hierarchical models.

Who Was Thomas Bayes?

"An Essay Towards Solving a Problem on the Doctrine of Chances." (1763)

$$P(A|B) = \frac{P(B|A\,P)(A)}{P(B)}$$

Bayesian statistics is based on an interpretation of Bayes' theorem.

Bayes Theorem

We want to flip the question from *"How likely are these data if a hypothesis were true?"* to *"How likely is the hypothesis now that we've seen these data?"*

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

$P(\theta|y) =$ Posterior: probability of the parameters given the data

$P(y|\theta)P =$ Likelihood: probability of observing the data given the parameters.

$P(\theta) =$ Prior: represents external knowledge about the parameters

In English, the posterior is proportional to the likelihood times the prior.

How to Choose a Likelihood and Prior?

- To choose a likelihood, you need to pick a probability distribution that matches your data, e.g.,
  - Is your data continuous and unbounded? Try a normal distribution.
  - Do you have count data that cannot be negative? Try a Poisson distribution.
- When choosing a prior distribution, we choose values that we believe capable of representing the reasonable range that the parameter can take, or come from a previous study.

Credible vs. Confidence Intervals

- Bayesian 95% credible interval: given the data and prior, there is a 95% probability the parameter lies in this interval.
- Frequentist 95% confidence interval: across repeated samples, 95% of the intervals produced by the method would cover the true value; it is not a 95% probability for this specific interval.
- Communication tip: credible intervals match how most people naturally talk about uncertainty.

Interpreting Bayesian GLM Coefficients

- Logistic GLM (binary): coefficients describe changes in log-odds; report odds ratios and changes in predicted probability for clarity.
- Poisson GLM (counts): coefficients describe changes in log rate; report rate ratios and predicted counts.
- Bayesian twist: report distributions and probabilities, e.g., “There’s a 92% probability the rate ratio exceeds 1.1.”

Advantages and Disadvantages

Advantages	Challenges
Full uncertainty quantification	Computationally intensive
Natural incorporation of prior knowledge	Choice of priors matters
Directly interpretable probability statements	Harder for very large datasets
Flexible for complex hierarchical models	Needs careful convergence diagnostics

How Does Bayesian Sampling Work?

- In Bayesian statistics, we want the whole distribution of plausible parameter values (the posterior), not just a single number estimate.
- We need a procedure that samples plausible values so we can compute the averages, intervals and predictions from those samples.
- Markov Chain Monte Carlo (MCMC) creates a sequence of guesses (a chain) for each parameter.
- Each new guess depends on the current one.
- If we run the guessing for long enough, the chain spends time in each region in proportion to how plausible It is under our model (likelihood x prior)
- The saved guesses are the posterior draws

How Does Bayesian Sampling Work?

- We pick an initial guess for the parameters (often random, and we use multiple chains with different initial guesses)
- From that point we begin to explore the landscape of plausibility defined by our likelihood and priors
- Early steps are just to find good regions and tune in the sampler – this is **warmup/burn-in** and we discard it.
- At each step, we propose a small move from the current guess. We take the current value and add a little random noise to form a step.
- If the step is too small, the chain will shuffle slowly and not fully cover the parameter space
- If the step is too big, the proposal will land in implausible places and get rejected

How Does Bayesian Sampling Work?

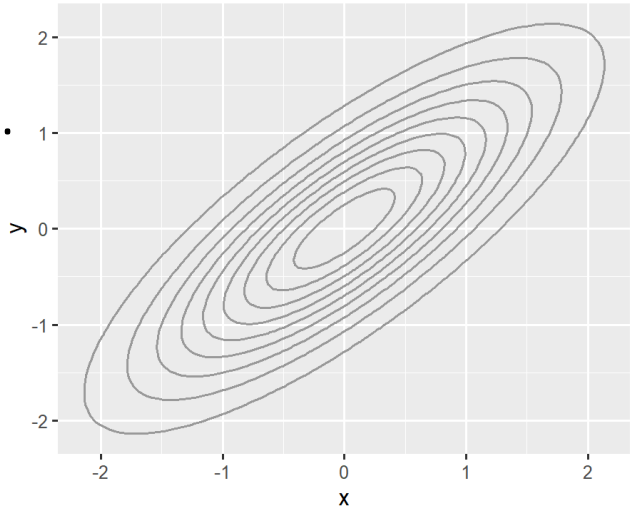
- We score the new guess by how well it explains the data and respects the prior. The rule of thumb is that if the proposal looks better than the current guess, accept it. If it looks worse, sometimes accept it anyway (with a small probability) so the chain is able to explore.

$$Accept\ with\ probability = \min\{1, \frac{posterior(new)}{posterior(current)}$$

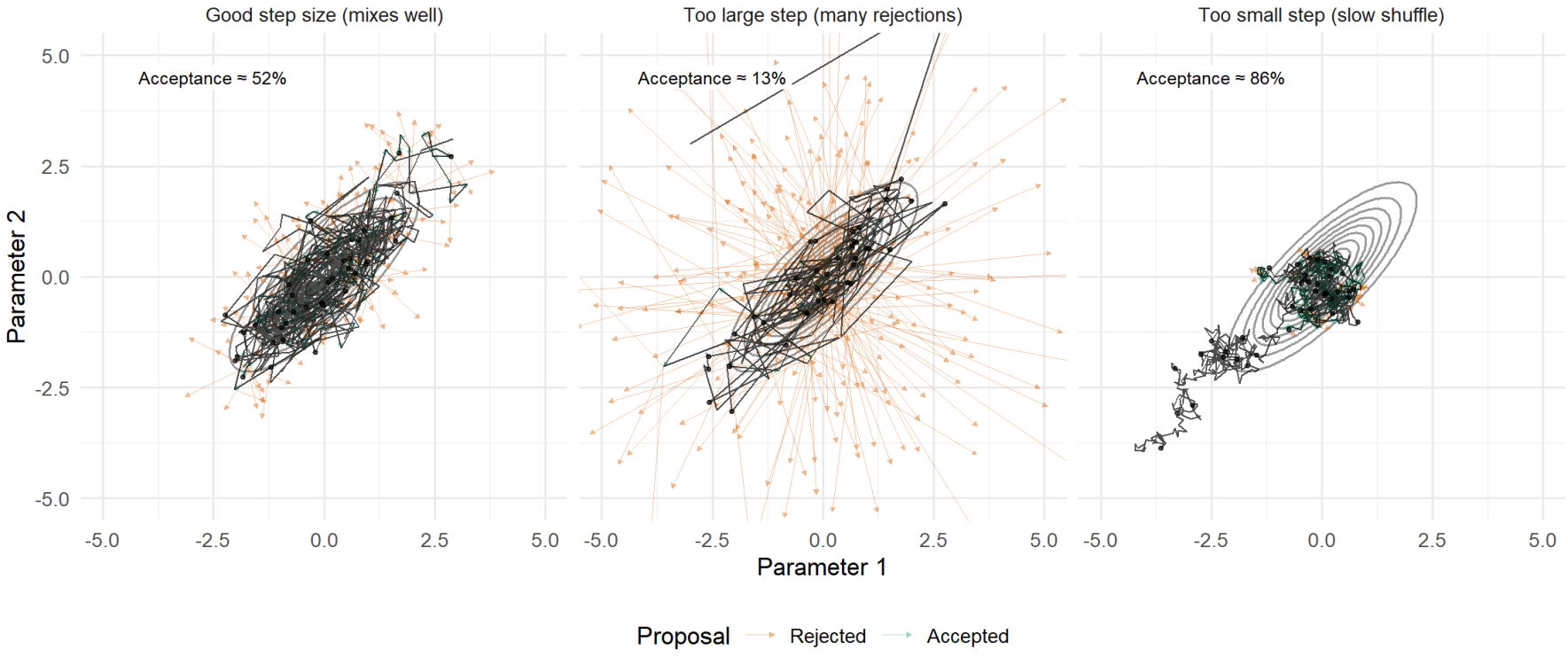
- Why accept 'worse' guesses? If we only ever accepted better guesses, we'd get stuck on a local peak. Occasional downhill moves lets the chain wander across the landscape and sample all plausible regions. The result is a sample that represents the whole posterior, not just the very top.
- We repeat the process of proposing a step, scoring it, and accepting/rejecting thousands of times. After warmup, the chain is considered settled and we keep the subsequent draws. The saved draws behave like random samples from the posterior.

How Does Bayesian Sampling Work?

- Each point is a possible pair of parameter values in a simple model with two parameters. A parameter value is plausible if it is probable under our assumptions and the data
- The contour lines represent the posterior landscape. Inner contours are more plausible, outer contours are less plausible. The centre is the most plausible combination of the two parameters
- We hope the sampler will find this area of high plausibility during warmup, and then wander across it, spending more time in the inner contours and some time in the outskirts. We do not want it to get stuck in any one place
- If we start multiple chains in different places, they should all end up exploring the whole area.



How Does Bayesian Sampling Work?



Frequentist Modelling Results

```
> summary(fit_glm)

Call:
glm(formula = y ~ x1 + x2, family = binomial(), data = dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6318     0.1301  -4.855 1.21e-06 ***
x1            0.9706     0.1123   8.645  < 2e-16 ***
x2           -0.5853     0.1952  -2.999  0.00271 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 747.65  on 599  degrees of freedom
Residual deviance: 646.30  on 597  degrees of freedom
AIC: 652.3

Number of Fisher Scoring iterations: 4
```

Bayesian Modelling Results

```
> summary(fit_brm)
 Family: bernoulli
  Links: mu = logit
Formula: y ~ x1 + x2
   Data: dat (Number of observations: 600)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.64      0.13    -0.89    -0.39 1.00     3703     3104
x1            0.97      0.11     0.76     1.20 1.00     3357     3101
x2           -0.59      0.20    -0.97    -0.20 1.00     3875     2925

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Prior Predictive Checks

- Prior predictive checks answer the question: “if my priors were correct, what kinds of data would I expect?”
- We simulate fake datasets from the model and priors before seeing the real outcomes. If the simulated data looks implausible (e.g., all successes or huge, impossible counts), its tells you your priors are too wide/tight/misaligned, and need to be fixed before fitting your model.
- You should run them before fitting your model, and re-run whenever you change priors, add predictors, switch links/families or change units/offsets
- If prior predictive outcomes are too extreme, slope priors should be tightened, or predictors centred. If they are too concentrated, priors can be loosened.
- The goal is weakly informative priors that rule out absurd data while staying compatible with domain knowledge

Posterior Predictive Checks

- Prior predictive checks answer the question: “if my model (likelihood + link + priors) were true, would it typically produce data like this?”
- We simulate datasets from the fitted model and compare those to the data that we observed. If the observed data look implausible under the posterior predictive, you’ve found model misfit (wrong family/link, missing predictors, lack of accounting for overdispersion etc.)
- You should run them after fitting your model (they assess fit after learning from the data), and re-run every time you change the model, adding variables, tweaking priors or changing family/link.
- Look for systematic mismatches: too many zeros, too-heavy tails, wrong mean/variance, under/over dispersion

Posterior Predictive Checks

- Binary/Logistic:
  - Overall success rate; group-wise rates across key predictors.
  - Calibration-style checks: predicted vs observed proportions.
- Counts/Poisson/NegBin:
  - Mean vs variance; zeros; right tail counts.
  - Rootograms to visualise discrete fit.
- General:
  - Quantiles (e.g., 0.05, 0.5, 0.95), ECDF overlays, residual-like checks on linear predictor.

Posterior Predictive Checks

- Common pp_check() options:
  - `pp_check(fit, type = "dens_overlay")`
  - `pp_check(fit, type = "hist")`
  - `pp_check(fit, type = "bars")`
  - `pp_check(fit, type = "rootogram")`
  - `pp_check(fit, type = "ecdf_overlay")`
- Pick visuals that make deviations obvious to a non-expert; label clearly.

Reproducibility and Efficiency

- Set seeds, save fitted models, and reuse compiled code to speed iteration.
- R code:

```
saveRDS(fit_brm_logit, "fit_brm_logit.rds");
saveRDS(fit_brm_pois, "fit_brm_pois.rds")
fit_brm_logit <- readRDS("fit_brm_logit.rds")  #
reuse later
```

$\hat{R}$ (R-hat/Gelman-Rubin)

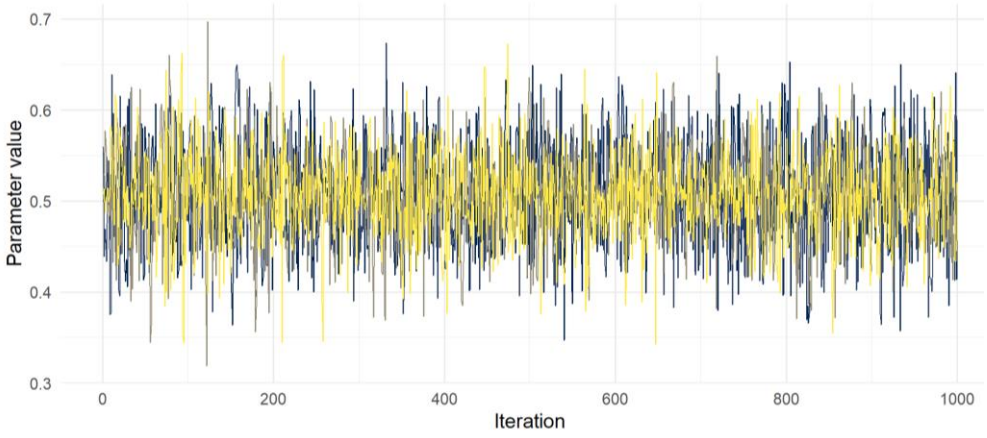
- $\hat{R}$ measures the between-chain vs. within-chain variance. It is used to check if chains are sampling the same posterior.
- Between-chain variance measures how far the chains lie from each other – ideally the chains should be exploring similar space, even though they start at different points.
- Within-chain variance measures the movement of draws within a single chain. This can be large just because the posterior itself is wide – a large within-chain variance does not mean it hasn't converged.
- means chains agree, whereas $\hat{R} > 1.01$ means the between-chain variance is larger than within-chain va riance, and the chains disagree on where the posterior is.
- $\hat{R}$ can be high due to too few iterations/warmup, weak priors, or unstandardised predictors

Effective Sample Size (ESS)

- MCMC samples are correlated (the draw at iteration I depends on the draw at iteration i-1) so N saved draws don't contain N units of fresh information.
- The ESS answers the question: "these N correlated draws carry the same information as how many independent draws?"
- The bigger the ESS, the tighter, more stable your summaries will be. A smaller ESS will lead to more noisy tables/intervals.
- Aim for at least 400 per parameter. 1,000 is good for stable intervals.
- To raise ESS try longer runtimes/warmup, standardising predictors, using different priors

Traceplots

- A traceplot is a time series of a parameter's sampled values across iterations, with a separate line for each chain (after warmup)
- After warmup, lines should hover in a stable band, lines from different chains should overlap, and lines should zig-zag and cross often.
- A thick, 'fuzzy caterpillar' appearance means good mixing
- If traceplots look bad, increase warmup, check prior distributions.



WAIC

- Widely Applicable Information Criterion (WAIC) approximates the model's expected predictive performance on new data, averaging over posterior uncertainty in the parameters.
- It outputs a single WAIC number. The smaller the WAIC value, the better the expected performance on new data.
- It can be thought of as the Bayesian analogue to AIC values, and is typically more appropriate for Bayesian fits, as it uses the posterior rather than just a point estimate (as AIC does) to approximate predictive performance.
- To use it, fit candidate models to the same data, and compare WAIC values, and prefer the model with the smaller WAIC value.

Hypotheses in brms

- R code:

```
hypothesis(fit, "b_x1 > 0")              # Pr(β_x1 > 0)
hypothesis(fit, "exp(b_x1) > 1.2")       # Pr(OR > 1.2)
```

- Report readable statements: 'There is a 91% probability the odds ratio exceeds 1.2.')

Example

- Consider an example where the count of bird species depends on habitat area.
- Y_i is the number of bird species observed at site i
- $Area_i$ is the habitat area for site i in hectares.
- The response variable Y_i follows a Poisson distribution

$$Y_i \sim Poisson(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 Area_i.$$

- Each additional hectare of habitat changes the expected count of bird species by a multiplicative factor of e^{β_1}

Coding Demo

Reporting Bayesian Results

- Priors: exactly what you used and why
- Computation diagnostics: brief assurance the fitting worked
- Posterior summaries: clear intervals and interpretable probabilities for effects of interest.
- Predictive checks: plots showing the model reproduces key features of the data.
- Sensitivity: note whether conclusions change under alternative reasonable priors.
- Avoid ‘significant/non-significant’ terminology

Common Misconceptions

- "Bayes is subjective." → Priors can be weak and documented; transparency beats hidden assumptions.
- "The prior dominates." → With reasonable priors, as n grows the data dominate.
- "It's too complex." → `brms`/`rstanarm` use familiar formula syntax; you focus on modelling, not sampler internals.
- "Using prior knowledge is cheating." → Science accumulates knowledge; ignoring it can be wasteful or misleading.