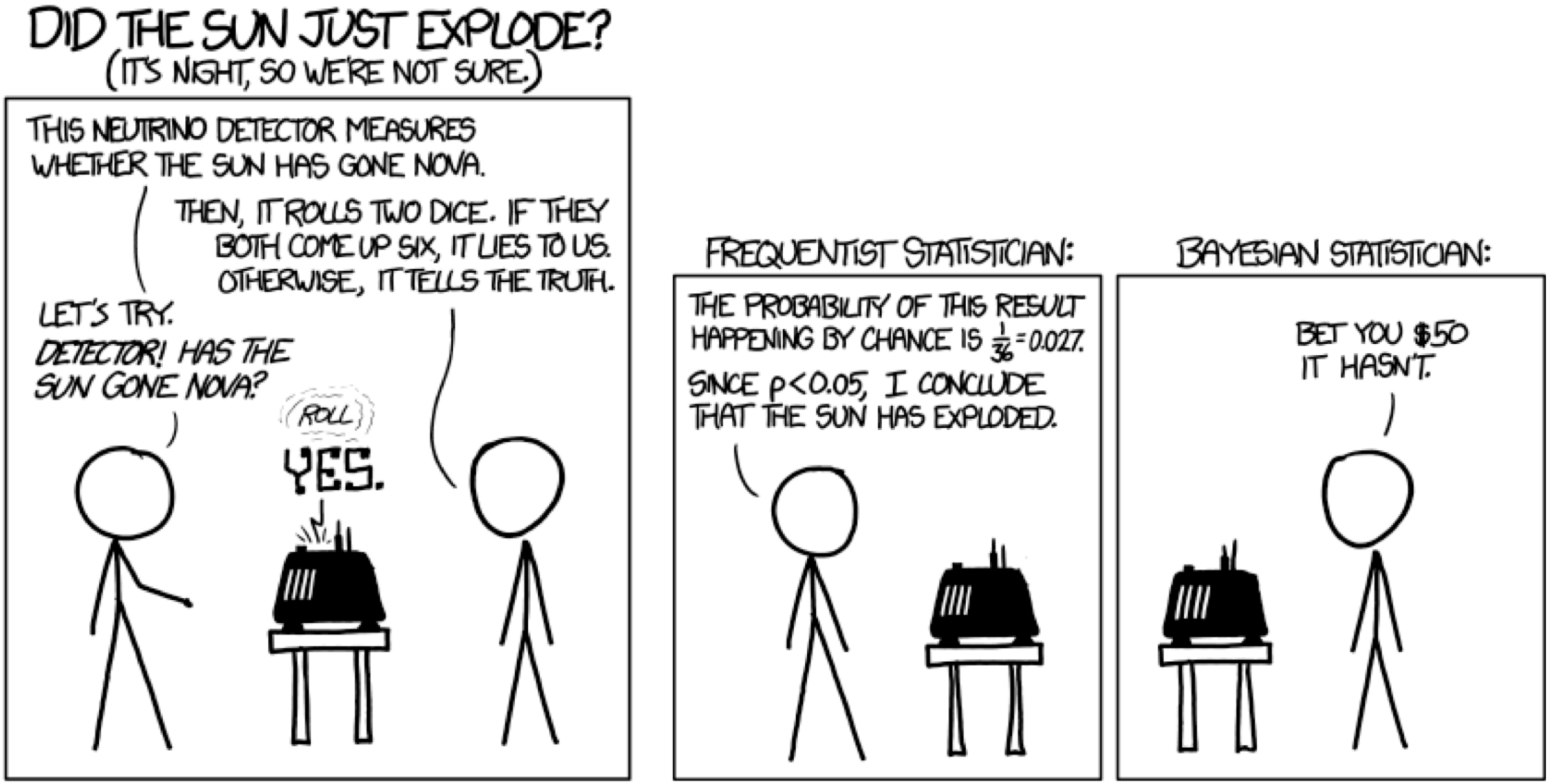
MIMM4750G
Bayesian inference



Likelihood

- Recall that likelihood is focusing on the probability as a function of the model parameters (hypothesis) given the data.
- For a given data set, we want to find the parameters that maximize the likelihood of our model.
- This is quite powerful!

Objections to maximum likelihood

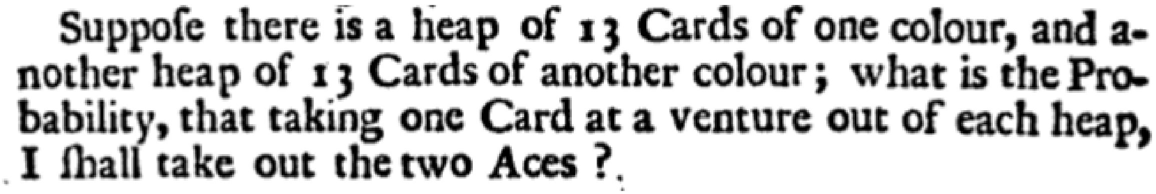
- The maximum likelihood estimate (MLE) is a single combination of parameter values.
- If the likelihood function is "rugged", then there may be many parameter values that are about as good as the MLE!

Being Bayesian

- A Bayesian would object to relying on a single estimate
- It is more robust to refer to the *distribution* of parameters that are supported by the data.
- A Bayesian would also object to the assumption that the experiment is completely objective.
- The design of an experiment is shaped by an investigator's subjective expectations about the outcome.

Rev. Thomas Bayes

- A Presbyterian minister in 17th century England.
- "An Essay towards solving a Problem in the Doctrine of Chances" was published two years after his death.
- Addressed a hypothetical gambling problem proposed by Abraham de Moivre:

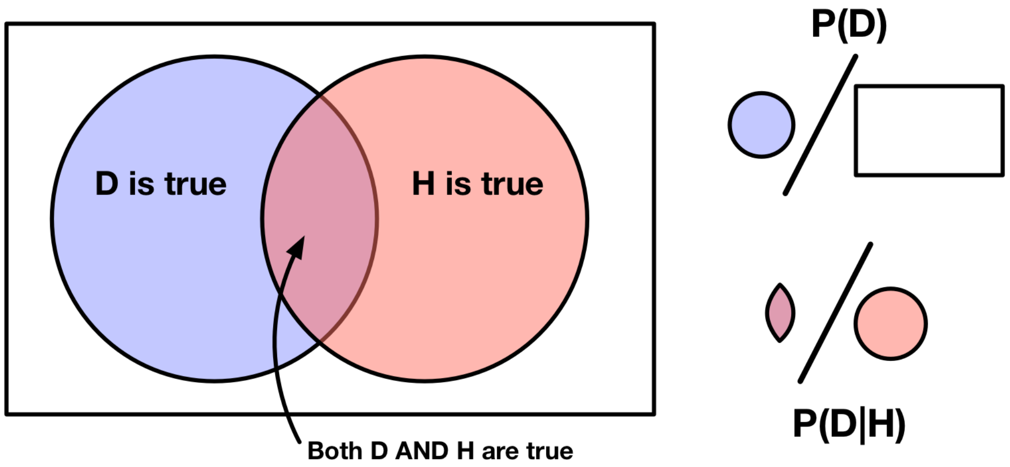


Suppoſe there is a heap of 13 Cards of one colour, and another heap of 13 Cards of another colour; what is the Probability, that taking one Card at a venture out of each heap, I ſhall take out the two Aces?

This may be a portrait of Thomas Bayes, but no one is really sure. We use it anyhow.

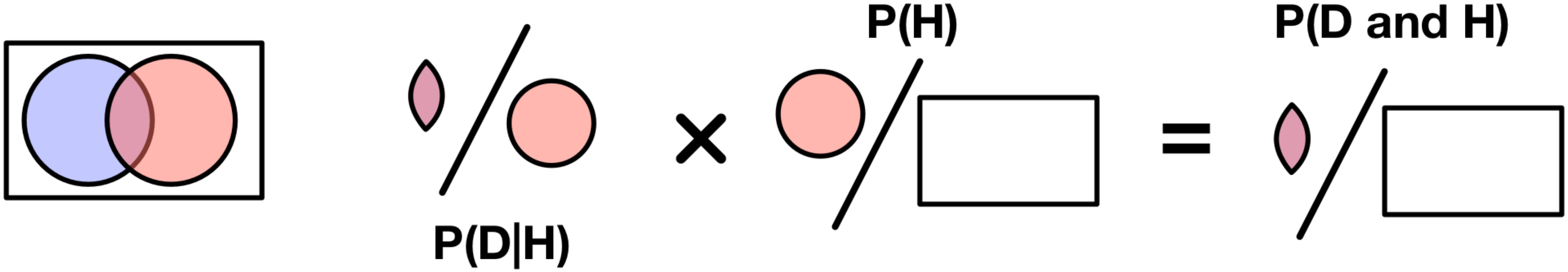
Conditional probability

- We are used to thinking about the probability of an outcome, $P(D)$.
- This implicitly involves some model (hypothesis), $P(D|H)$.
- We say that $P(D|H)$ is "the probability of the data, conditional on the hypothesis being true".



Joint probability

- We can calculate the **joint probability** that both D and H are true:



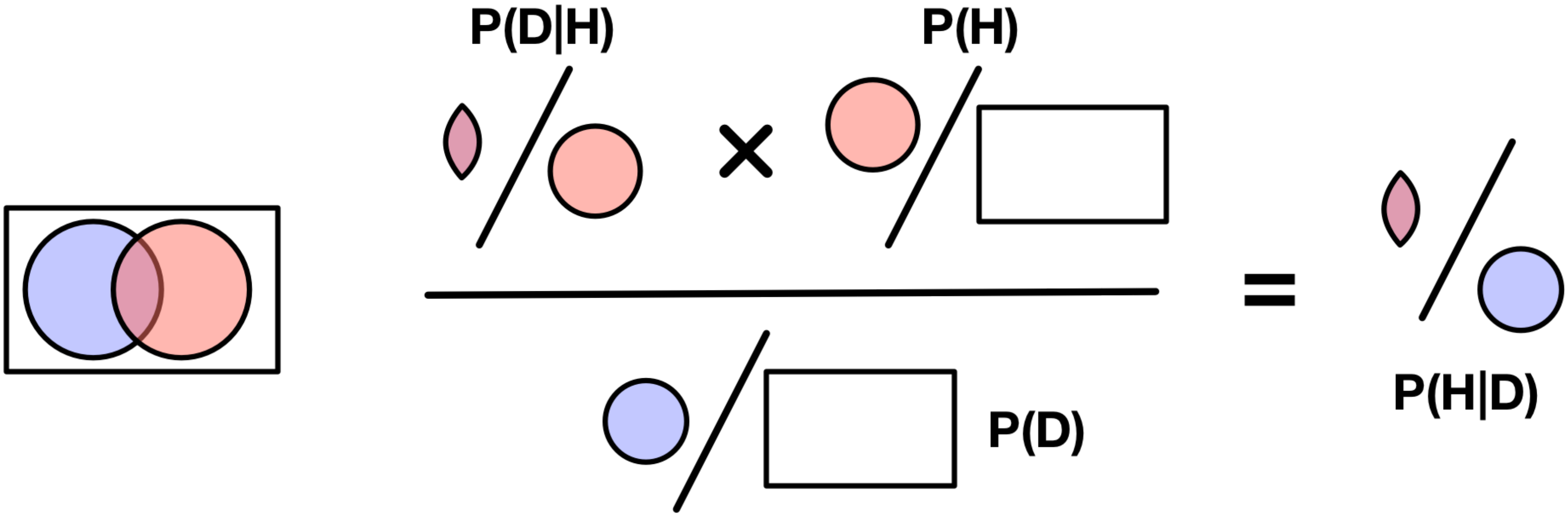
$$P(D, H) = P(D|H) \times P(H)$$

- It's perfectly valid to swap D and H!

$$P(D, H) = P(H|D) \times P(D)$$

Bayes' theorem

- This leads to something outrageous and wonderful:

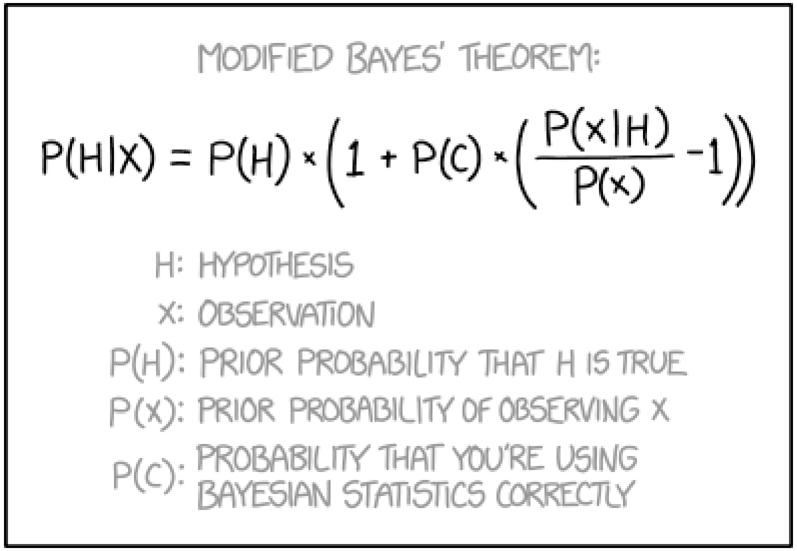


Belief

- This formula has some strange quantities.
- $P(D|H)$ is the likelihood.
- $P(D)$ is the probability of the data. *Weird*.
- $P(H)$ is the probability of the hypothesis without any data.
- If we have no data, then we can only work with our *prior belief*.
- $P(H|D)$ is then our updated belief *after we have seen the data*. It is the posterior belief.

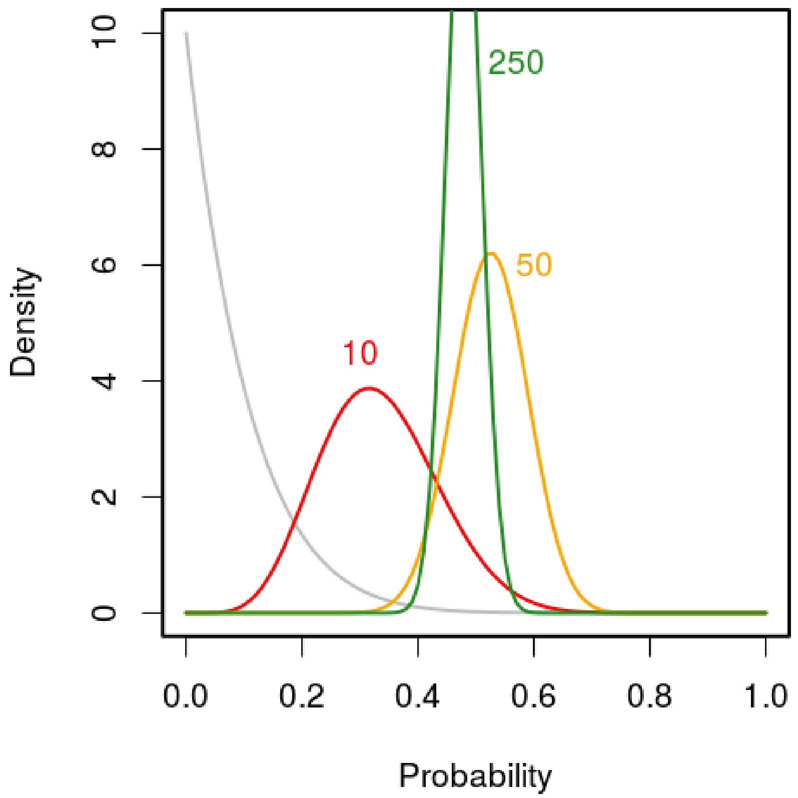
Reasons why people don't like Bayes' theorem

- "Belief doesn't seem scientific."
- "How am I supposed to decide what my prior belief is?"
- "The prior is biasing your study."
- Too much weird math.



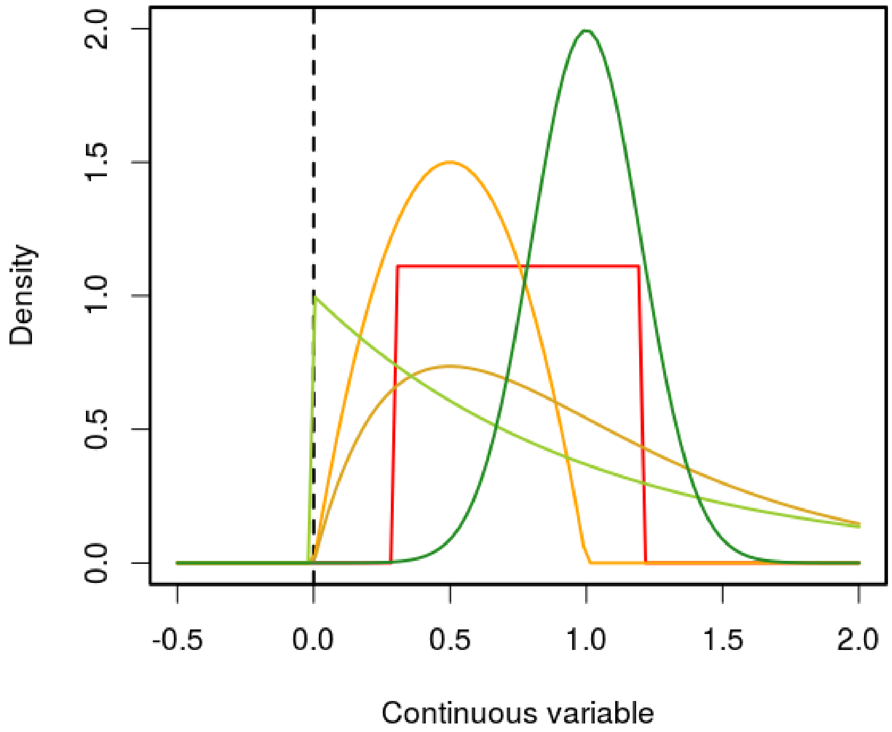
Embrace the prior

- Priors are natural: I have an expectation that when I toss a coin toss, it will come up heads about 50% of the time.
- Priors are flexible (unless you have very little data).
- (right) Updating prior with 10, 50 and 250 coin tosses given true probability is 50%.



Choosing your prior

- There are many probability distributions that we can use as priors.
- Uniform (red)
- Beta (0,1) (orange)
- Gamma (0,∞) (yellow)
- Exponential (0,∞) (light green)
- Gaussian ($-\infty$, ∞) (dark green)



Getting rid of P(D)

- We could calculate $P(D)$ exactly by integrating $P(D|H)$ over *all possible hypotheses*:

$$P(D) = \int_H P(D|H)P(H)$$

- It is often not possible to solve for this integral.
- We need to take a different approach...

What do we *really* want?

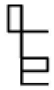
- We *really* want to know what the posterior distribution $P(H|D)$ looks like.
- For example, what is the *mean* or *median* value? (What is our best guess about the true value of H?)
- Usually, $P(H|D)$ cannot be written down as a mathematical formula.
- The next best thing would be a random sample of values H from $P(H|D)$.

Monte Carlo

- Stanislaw Ulam was a Polish physicist who, in 1946, was playing solitaire while recovering from brain surgery.
- He reasoned that it would be easier to estimate the probability of winning by playing many times (simulation) than calculating the exact chance.
- This simulation-based approach was dubbed the "Monte Carlo method" after the casino.
- The method later became used in the Manhattan project.

Random walks

- Suppose you exit the classroom and after every 10 steps, you flip a coin twice.
- HH, face forward
- HT, turn left
- TH, turn right
- TT, turn around



Markov chain Monte Carlo

- A random walk is basically a simulation. If we use a random walk to solve a problem, we are using a *Monte Carlo method*.
- Remember a *Markov chain* is a random process where the probability of the next event depends only on the current state (like Snakes and Ladders).
- Markov chain Monte Carlo (MCMC) is a powerful method to solve problems in a Bayesian framework.

Metropolis-Hastings sampling

- Remember that $P(D)$ is a difficult integral to deal with.
- What if we consider the ratio of $P(H|D)$ for two hypotheses, H and H'? Then $P(D)$ cancels out!
- M-H sampling is a random walk over the space of model parameters (H).
- We *propose* a new set of parameters H', and then decide whether to accept this proposal.

Metropolis-Hastings sampling

- Our random walk is controlled by this ratio:

$$R = \frac{Q(H'|H)}{Q(H|H')} \times \frac{P(D|H')P(H')}{P(D|H)P(H)}$$

  where $Q(y|x)$ is the probability of proposing y when you are at x.

- Remember that $P(H|D) \propto P(D|H)P(H)$.

Metropolis-Hastings sampling

- It turns out that if you follow these rules:

  1. Always accept the proposed H' if $R \geq 1$.
  2. If $R < 1$, accept H' anyways with probability R.

     This means we take a "step down"!

  3. Otherwise, stay where we are with H.

- then the amount of time our random walk spends in H will be proportional to the posterior probability $P(H|D)$.

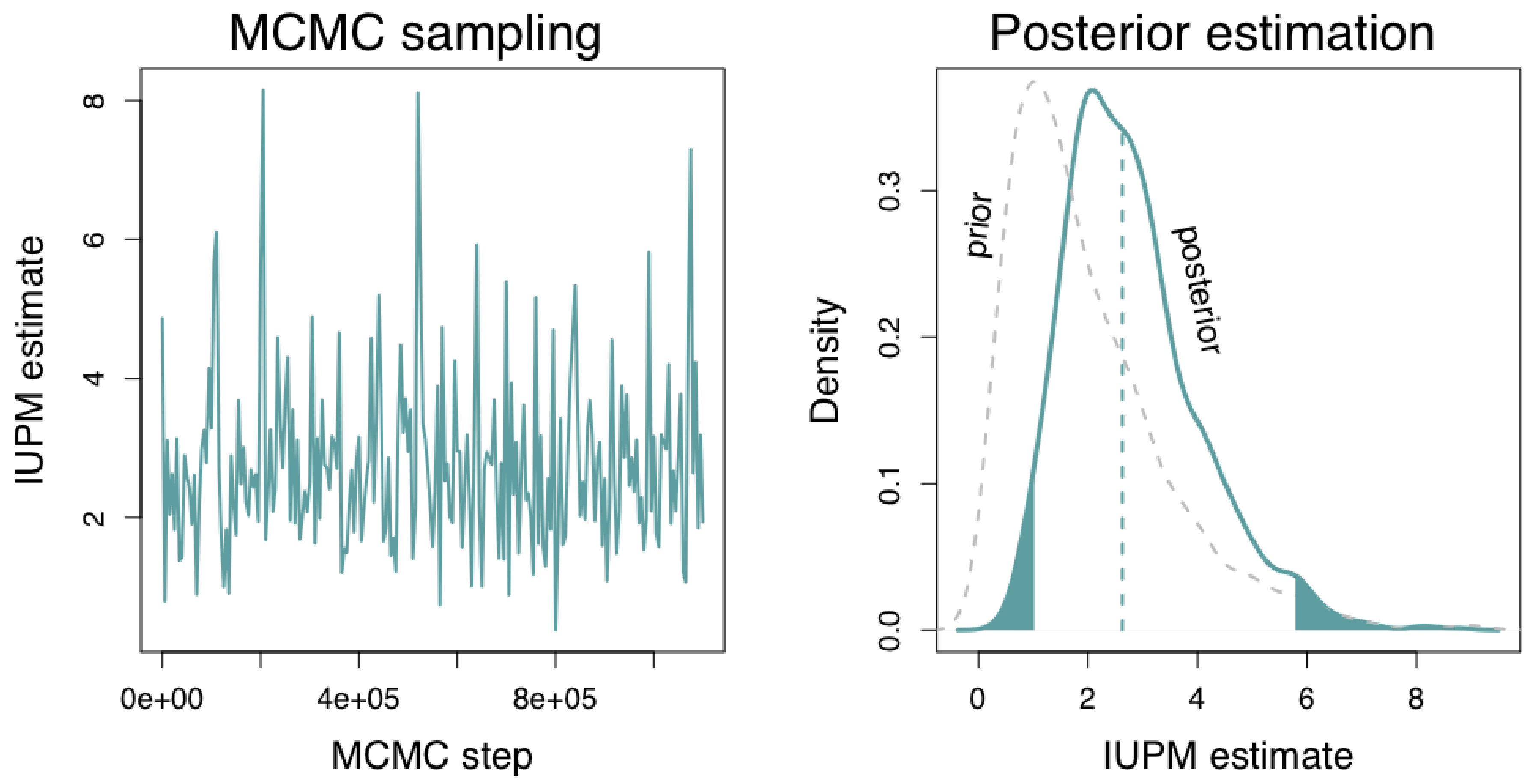


Figure from Poon *et al.* (2018) Retrovirology 15:47.

Convergence

- MCMC is an "auto-correlated" process - the current state will always be similar to the previous state.
- This is *efficient* because we don't waste time sampling states (parameter values) that are silly.
- This is *not efficient* because a random walk is slow to explore parameter space.
- When a random walk has gone long enough, it should eventually "converge" to the posterior distribution.

Burn-in and thinning

- What if the random walk starts in an awful part of parameter space?
- We don't want this to affect the sample, so we throw out the first part of the walk (chain).
- To reduce auto-correlation the sample, we only keep a small number of samples taken from equal intervals along the chain (thinning).