Goals

- Joint, conditional, and marginal probability
- Bayes' rule and Bayesian updating
- Markov Chain Monte Carlo

These three topics are essential to understanding modern Bayesian statistics.

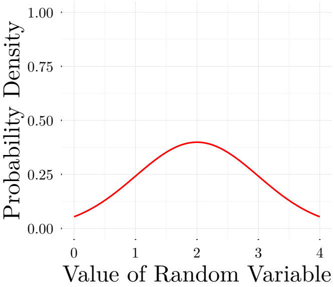
Goals

Advantages of Bayesian approach:

- Incorporates prior information in a formal, mathematically sound framework
- Fits arbitrarily complex models, especially multilevel models
- Works directly with probabilities, straightforward interpretation

Probability Review

A **probability density function (PDF)** describes distribution of a random variable (RV).



A PDF is **parametric** if its shape is governed by parameters (ie. mean, scale, shape, etc.) Notation, ie. normal PDF:

$$\Pr(X) = \frac{1}{2\sigma\sqrt{\pi}} \exp\{-\frac{(X-\mu)^2}{2\sigma^2}\}$$

Probability Review

The PDF gives the probability that random variable X falls within an infinitesimal interval: [X, X + really small number].

Hence, the PDF gives the probabilities of all mutually exclusive outcomes of random event X.

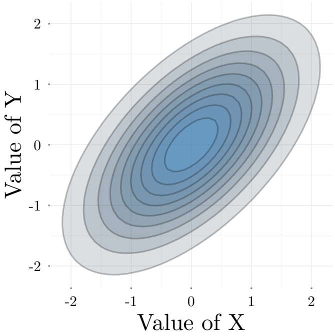
The sum of the probabilities of all mutually exclusive outcomes **must equal 1**. This is the area under the PDF curve:

The integral of any PDF **must equal 1**.

Joint Probability Distribution

With >1 RVs, we have a **joint distribution function**.

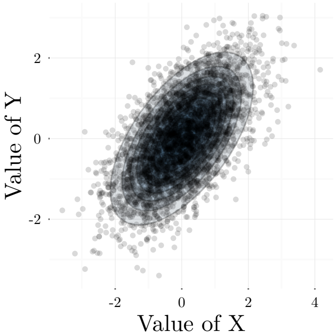
For 2 RVs, the joint distribution gives the probability of a pair of values (X, Y).



Notation: $\Pr(X, Y)$

Joint Probability Distribution

Sample data from this bivariate density ...
and we get more points where density is highest.

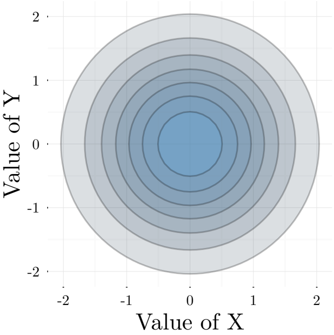


There is **dependency** between X and Y in this joint distribution:

$$\Pr(\text{large } X, \text{large } Y) > \Pr(\text{small } X, \text{large } Y)$$

Joint Probability Distribution

Contrast this with an indepedent X and Y.



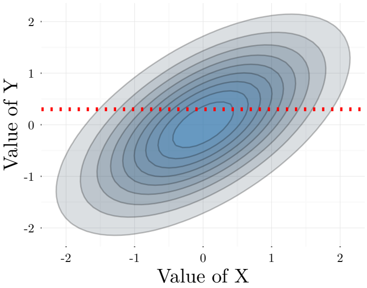
The probability of two independent events is the **product of the probabilities** of each of these events:

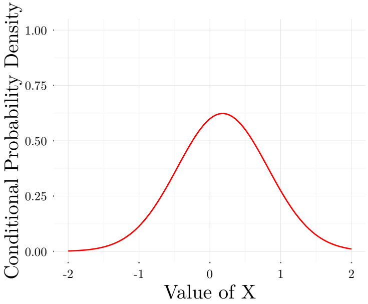
$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y) \text{ if } X \perp\!\!\!\perp Y$$

Conditional Probability Distribution

To characterize depedency, use the distribution of X conditional on a value of Y.

This is the probability that X takes on a value, given that Y has already taken on a value ($Y = y$).

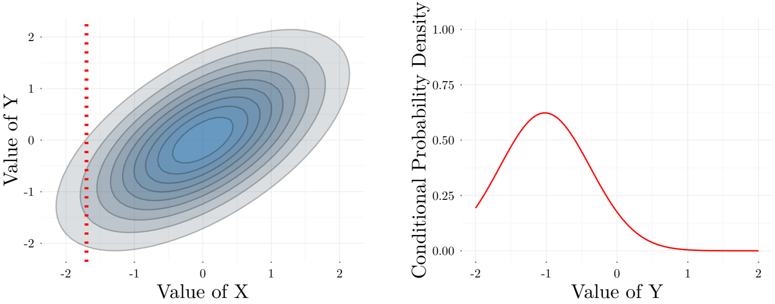




This is proportional to a slice of the joint density where Y equals a fixed value.

Conditional Probability Distribution

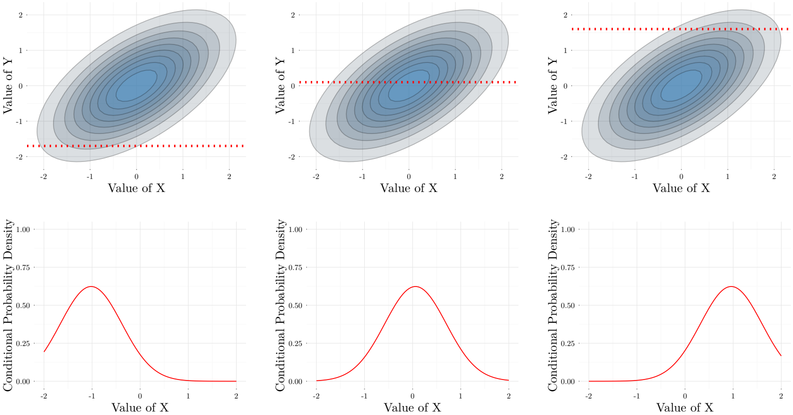
And vis versa: a slice where Y is conditional on a fixed value of X.



Notation: $\Pr(X|\,Y = y)$ or $\Pr(X|\,Y)$

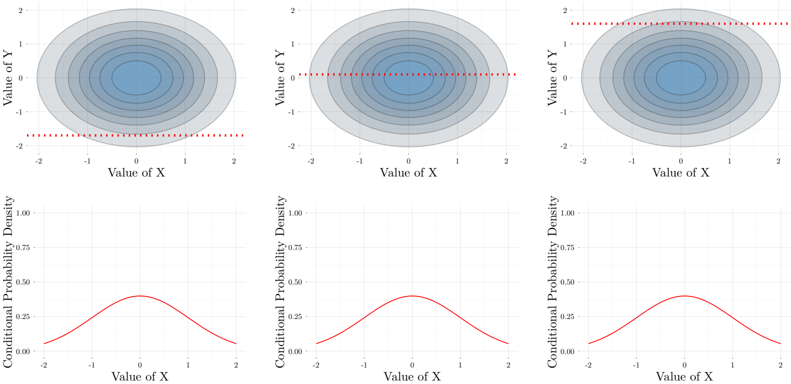
Conditional Probability Distribution

When **joint distribution is dependent**, $\Pr(Y|X)$ changes with values of X.



Conditional Probability Distribution

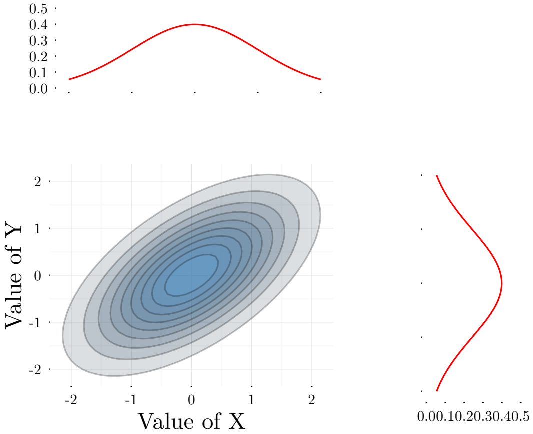
When joint distribution is independent, the $\mathrm{Pr(X|Y)}$ will be invariant as Y changes.



Hence $\Pr(Y|X) = \Pr(Y)$ if $Y \perp\!\!\!\perp X$.

Marginal Probability Distribution

In our joint density plot, the **marginal probability** is the probability density on the ‘edges’.



Identities

For X, the marginal distribution is constructed by summing all the 'slices' of the joint density at all values of Y.

Thus the marginal distribution is an integral:

$$\Pr(X) = \int \Pr(X|Y)\Pr(Y)dY$$

The probability that X equals a value conditinal on a value of Y, averaged across the probability density of Y.

Identities

The joint distribution gives the probability that X equals a value $(X = x)$ and Y equals a value $(Y = y)$:

- if X is independent from Y:

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)$$

- if X is dependent on Y:

$$\Pr(X, Y) = \Pr(X|Y) \cdot \Pr(Y)$$

The joint probablity of $(X = x, Y = y)$ can be understood as the probability that two events happen simultaneously: $(Y = y)$ and $(X = x)$.

Identities

The conditional distribution is the probability density of X given that Y already equals a value.

Think of this as the probability of values of X, given that $Y = y$ already happened.

$$\Pr(X|Y = y) = \Pr(X, Y)/\Pr(Y)$$

If X is independent from Y:

$$\Pr(X|Y = y) = \Pr(X)\Pr(Y)/\Pr(Y) = \Pr(X)$$

Conditional, marginal, and joint densities all integrate to 1.

Parameters and Data

Broadly, our goal in statistics is to make inference on the underlying process (parameters) which could have generated our observed data.

Example: we collect data on the growth rate, of a plant species transplanted to a novel environment.

- The observed growth rates of individual seedlings are data.
- The average growth rate is an unknown parameter to estimate.

What we're interested in: the probability distribution of the average growth rate conditional on the data.

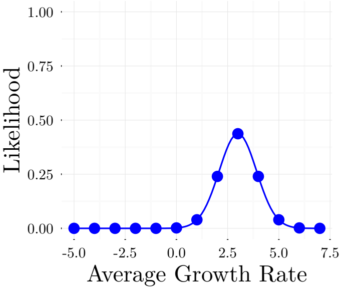
$$\Pr(\text{average growth rate}|\text{observed growth rates})$$

Parameters and Data

If assume a probability distribution for our data (say a normal distribution), we can easily calculate:

$$\Pr(\text{observed growth rates}|\text{average growth rate})$$

For all potential values of the average growth rate.



This is called the **likelihood of the parameter**, given the data.

Parameters and Data

Likelihood is a function of probability, but is not itself a probability. The distinction is that the data are treated as a random variable, while the parameter is not.

We care about likelihood insofar as it informs us about the probability of various values of a parameter.

We are not interested in the probability of the data given various values of the parameter.
(ie. how probable is an observed growth rate of 3, given a average growth rate of μ?).

We are interested in the probability of the parameter given our data
(ie. how probable is an average growth rate of μ, given that we observed a growth rate of 3?).

Bayes Rule

Bayes' rule takes a conditional probability and inverts it.

$$\Pr(Y|X) \Rightarrow \Pr(X|Y)$$

Useful to convert Pr(data|parameters) into Pr(parameters|data).

Bayes' rule (in the context of data):

$$\Pr(\text{parameter}|\text{data}) = \frac{\Pr(\text{data}|\text{parameter})\Pr(\text{parameter})}{\Pr(\text{data})}$$

i.e., if we observe a single growth rate of $y = 3$, the probability of an average growth rate of $\mu = 10$ conditional on this data:

- the marginal probability that the average growth rate is 10,
- the conditional probability that observed growth rate is 3, given that the average growth rate is 10.
- inverse of the marginal probability that the observed growth rate is 3.

Bayes Rule

$$\Pr(\text{parameter}|\text{data}) = \frac{\Pr(\text{data}|\text{parameter})\Pr(\text{parameter})}{\Pr(\text{data})}$$

Terminology:

- $\Pr(\text{parameter}|\text{data})$ is the **posterior**.
- $\Pr(\text{data}|\text{parameter})$ is the **likelihood**.
- $\Pr(\text{parameter})$ is the **prior**.
- $\Pr(\text{data})$ is the **normalizing constant**.

Bayes Rule: Numerator

$$\Pr(\text{parameter}|\text{data}) \propto \Pr(\text{data}|\text{parameter})\Pr(\text{parameter})$$

The prior represents our **previous knowledge** (or previous belief) about the feasible values of a parameter.

The probability of the data conditional on the parameter (the likelihood), is our **evidence** for various values of the parameter.

Think of the likelihood as a weight:
the posterior is ***proportional*** our prior belief is *weighted* by our evidence, for any given value of the parameter.

Bayes Rule: Numerator

For example,

We assign a high prior probability that the average growth rate $(\mu) = 10$:

$$\Pr(\mu = 10) = 0.27$$

If the observed growth rate $(y) = 3$ is improbable, given $\mu = 10$:

$$\Pr(y = 3|\mu = 10) = 0.000004$$

Then the posterior probability of $\mu = 10$ is very low relative to the prior:

$$\Pr(\mu = 10|y = 3) = 0.0000007$$

Note that the posterior is not equal to the product of likelihood and prior: this is because of the denominator in Bayes' rule.

Bayes Rule: Numerator

With a different prior:

We assign a high prior probability that our average growth rate $\mu = 2.7$:

$$\Pr(\mu = 2.7) = 0.27$$

The observed growth rate $y = 3$ is quite probable given this average growth rate:

$$\Pr(y = 3|\mu = 2.7) = 0.26$$

In this case the posterior probability of $\mu = 2.7$ is high relative to the prior:

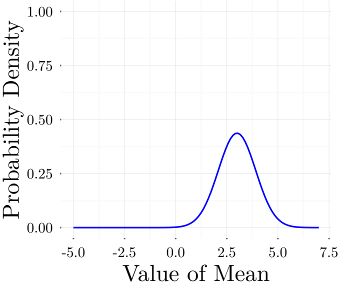
$$\Pr(\mu = 2.7|y = 3) = 0.66$$

The more data we have, the stronger the weight.
The greater the uncertainty (variance) in the prior, the weaker the prior.

Bayes Rule: Numerator

The likelihood (information in the data) thus ‘updates’ the prior information: this is **Bayesian Updating**.

Imagine we have a single observation (y = 3), from a distribution with a known variance.

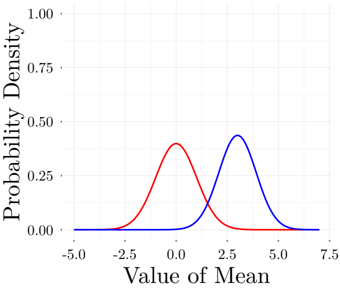


The probability of the data point given various values of the mean μ, is shown in blue. This is the likelihood.

Bayes Rule: Numerator

If we set a prior that represents an equivalent amount of information:

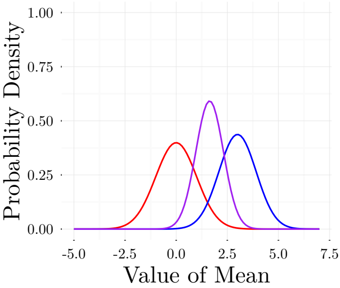
$$\mu \sim \mathcal{N}(0, 1)$$



This prior probability of the mean, is shown in red.

Bayes Rule: Numerator

The posterior (purple) is intermediate between the two.

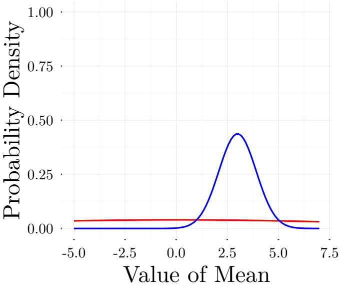


Thus, the data update the prior.

Bayes Rule: Numerator

Same likelihood with an uninformative prior on the mean (μ):

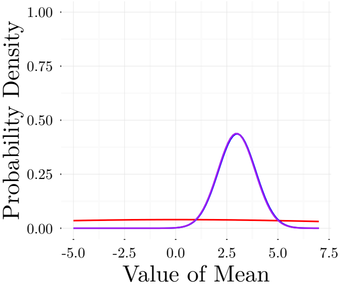
$$\mu \sim \mathcal{N}(0, 100)$$



Again, the prior probability is in red. Notice it is practically uniform.

Bayes Rule: Numerator

The resulting posterior is identical to the likelihood.



As the variance of the prior increases, it loses strength to influence the posterior.

Bayes Rule: Denominator

$$\Pr(\text{parameter}|\text{data}) = \frac{\Pr(\text{data}|\text{parameter})\Pr(\text{parameter})}{\Pr(\text{data})}$$

What is the denominator in Bayes' rule? This where the calculus voodoo comes into play.

Using our definition of marginal probability:

$$\Pr(\text{data}) = \int \Pr(\text{data}|\text{parameter})\Pr(\text{parameter})\, d\text{parameter}$$

The denominator is just the integral of the numerator!

Bayes Rule: Denominator

How does this make sense in any way shape or form?

The numerator is not a valid probability distribution: it is a product of probability distributions.

The numerator does *not* integrate to 1.

The numerator is a function of the parameters. Dividing by the integral normalizes this function, to integrate to 1.

The sole purpose of the denominator in Bayes' rule is to make the product of the likelihood and prior, into a valid probability distribution.

We don't care about the value of the denominator, but we **need** to calculate it.

Markov Chain Monte Carlo

Most of the time, the denominator is **difficult to calculate**.

This is because it is a high-dimensional integral; for example in single-level hierarchical normal model model:

$$\Pr(\mu, \theta_j, \tau^2, \sigma^2 | y_{ij}) = \frac{\Pr(y_{ij}|\theta_j, \sigma^2)\Pr(\theta_j|\mu, \tau^2)\Pr(\mu, \tau^2, \sigma^2)}{\Pr(y_{ij})}$$

The denominator in integral form is:

$$Pr(\mu, \theta_j, \tau^2, \sigma^2 | y_{ij}) = \int_{\theta_j}\int_{\mu}\int_{\tau^2}\int_{\sigma^2} Pr(y_{ij}|\theta_j, \sigma^2)Pr(\theta_j|\mu, \tau^2)Pr(\mu, \tau^2, \sigma^2)\, d\sigma^2\, d\tau^2\, d\mu$$

Markov Chain Monte Carlo

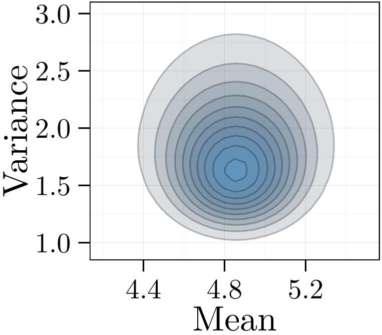
Markov Chain Monte Carlo (MCMC) is a clever numerical technique to simulate from the posterior distribution, without explicitly solving this integral.

To do so, we only need to know the conditional distributions for each parameter (which are unidimensional).

Markov Chain Monte Carlo

An example: have 34 datapoints we assume a normal distribution, and want to estimate the mean and variance.

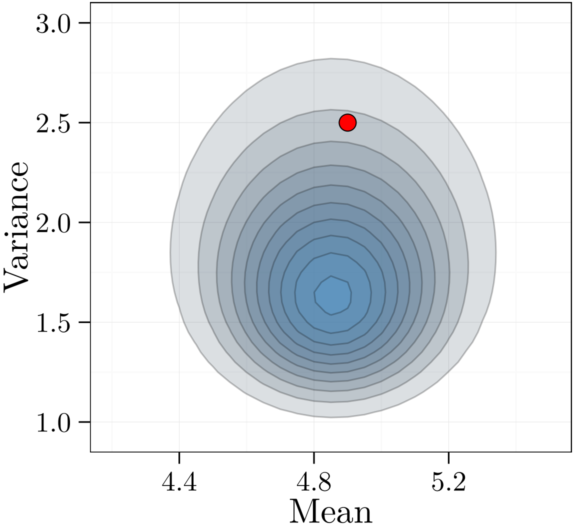
In this (example) case, we have an **analytical solution to the posterior**:



But we'll run an MCMC algorithm anyway to see how it works.

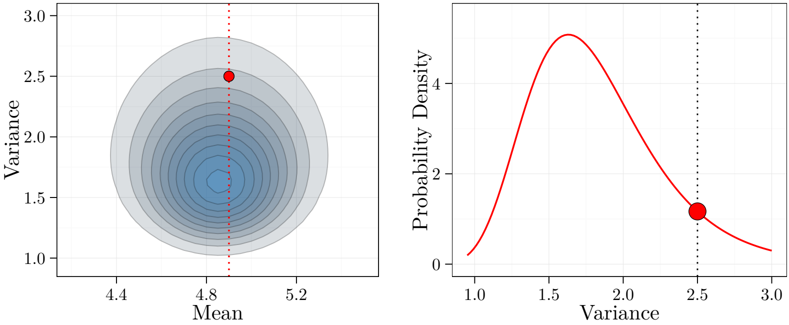
Markov Chain Monte Carlo: Initial Values

First we pick a reasonable starting value ($\mu = 4.8, \sigma^2 = 2.5$).



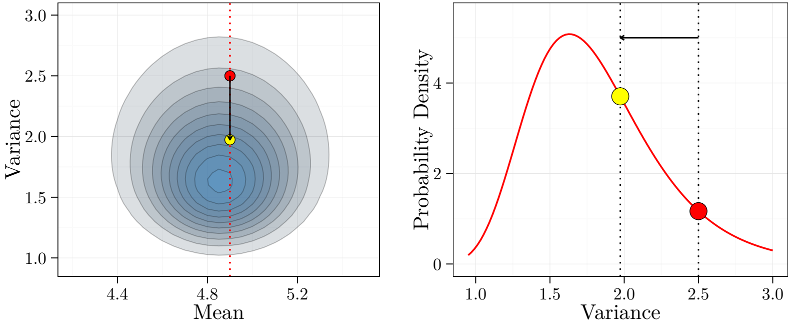
Markov Chain Monte Carlo: Conditional Distribution of Variance

Figure out conditional distribution for the variance (or proportional function), given the starting value of the mean.



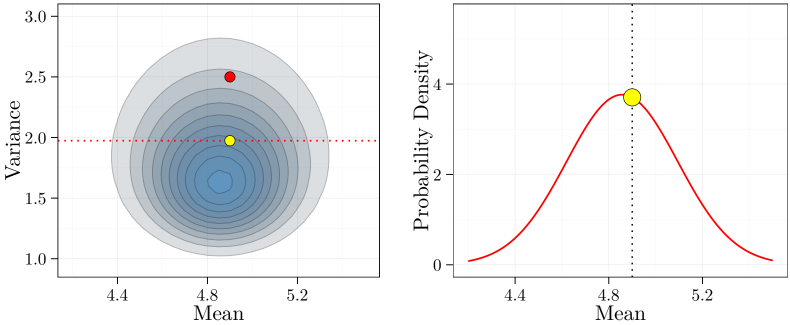
Markov Chain Monte Carlo: Update Variance

Simulate a new value of the variance from this conditional distribution.



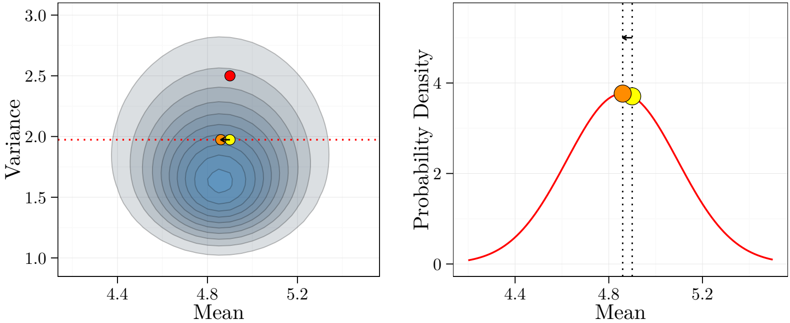
Markov Chain Monte Carlo: Conditional Distribution of Mean

Figure out the conditional distribution for the mean, given the new value of the variance.



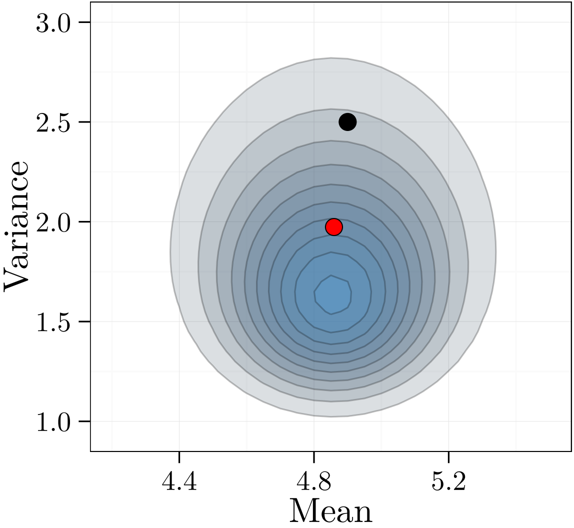
Markov Chain Monte Carlo: Update Mean

Simulate a new value of the mean from this conditional distribution.



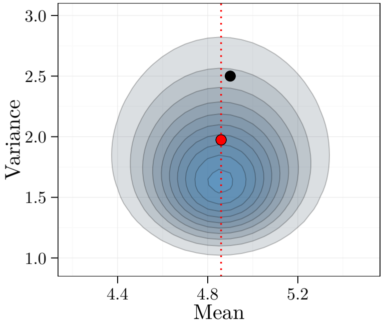
Markov Chain Monte Carlo: Iteration 1

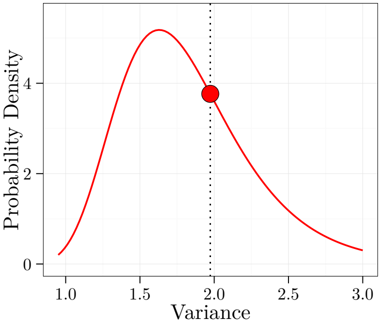
Now we have a new sample from the joint posterior distribution.



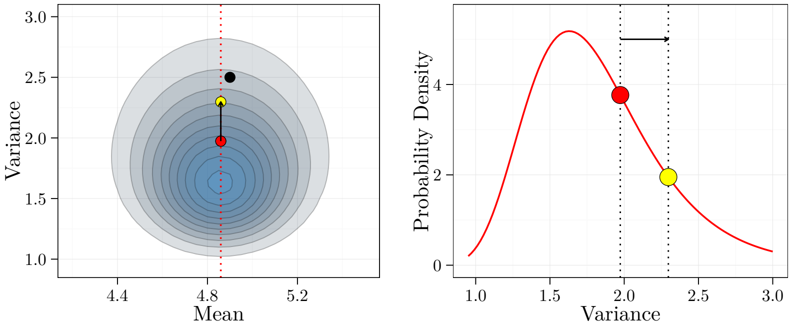
Markov Chain Monte Carlo: Conditional Distribution of Variance

Repeat ...

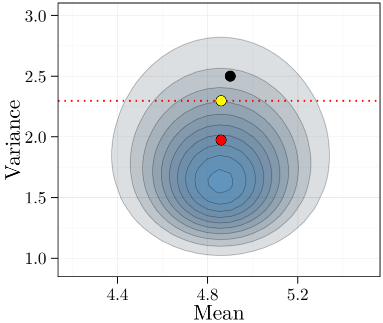


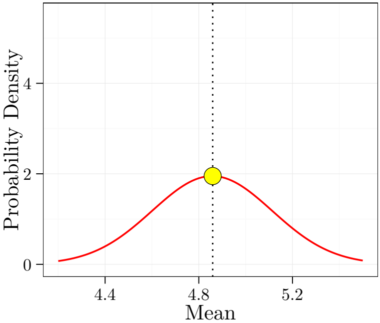


Markov Chain Monte Carlo: Update Variance

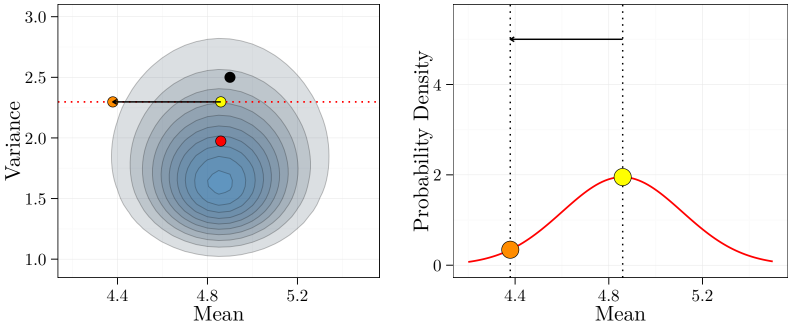


Markov Chain Monte Carlo: Conditional Distribution of Mean

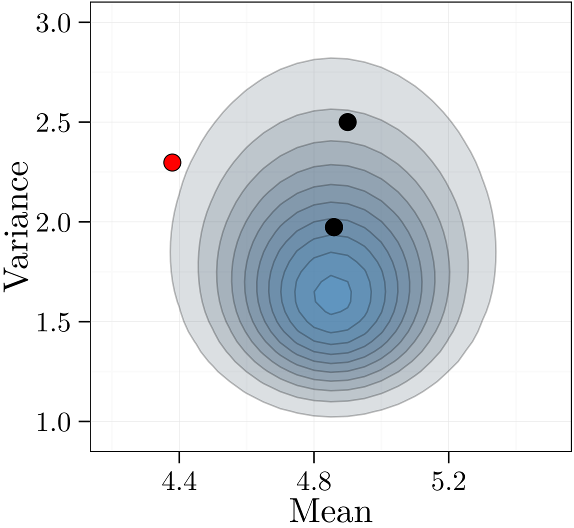




Markov Chain Monte Carlo: Update Mean

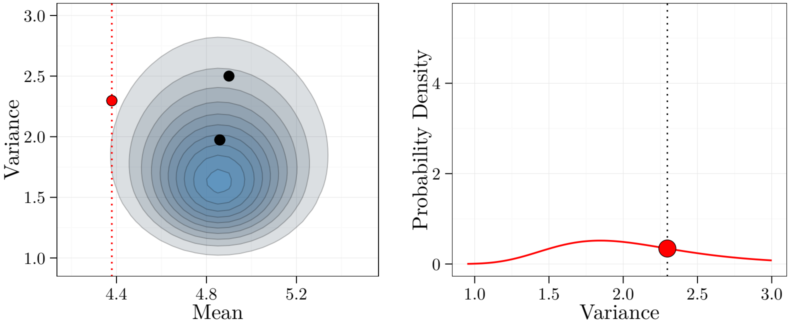


Markov Chain Monte Carlo: Iteration 2

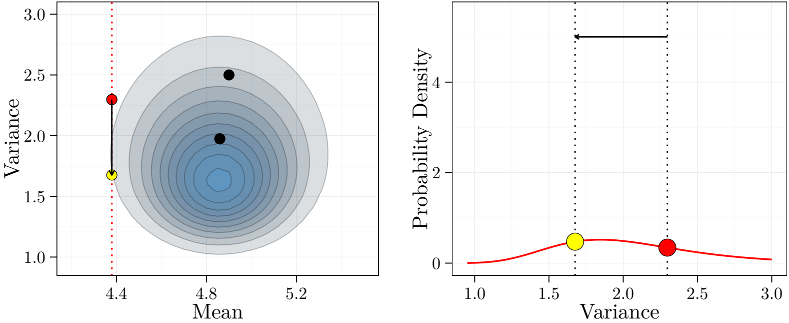


Markov Chain Monte Carlo: Conditional Distribution of Variance

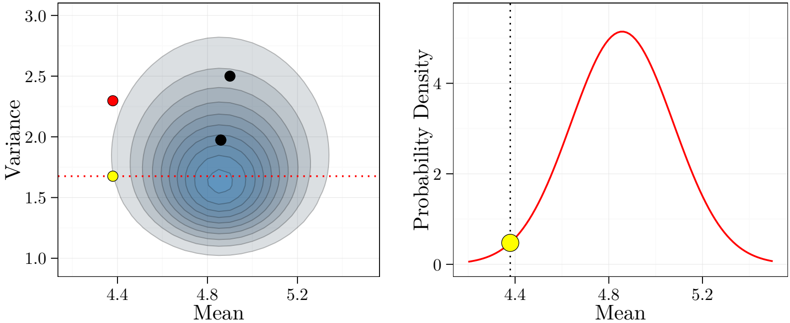
Repeat ...



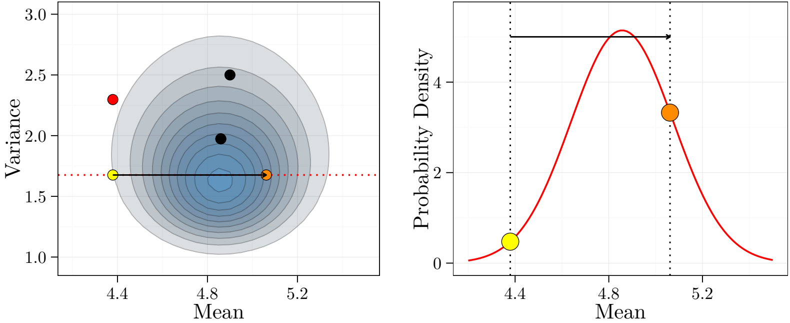
Markov Chain Monte Carlo: Update Variance



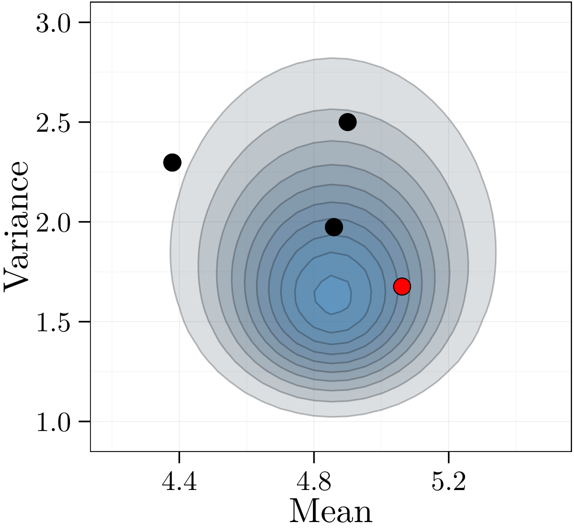
Markov Chain Monte Carlo: Conditional Distribution of Mean



Markov Chain Monte Carlo: Update Mean

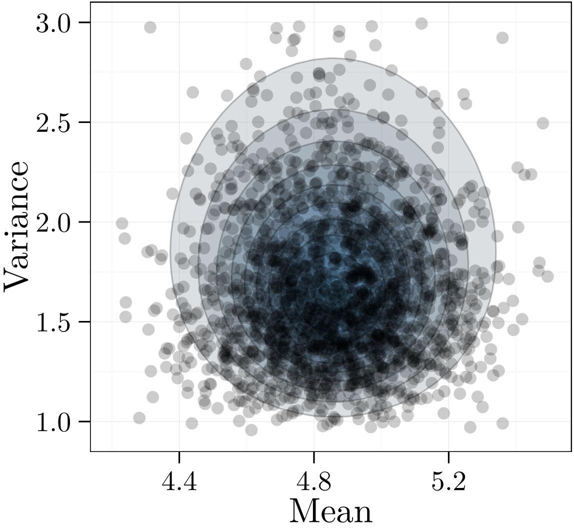


Markov Chain Monte Carlo: Iteration 3



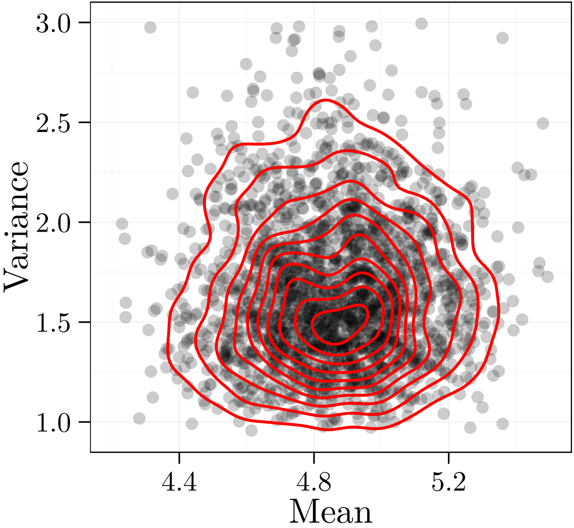
Markov Chain Monte Carlo: Result

Do this many, many times.



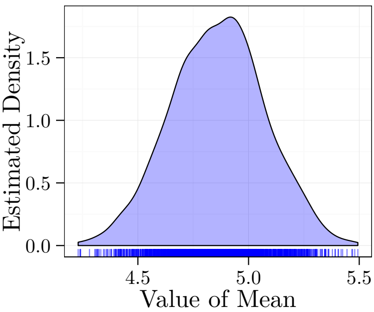
Markov Chain Monte Carlo: Result

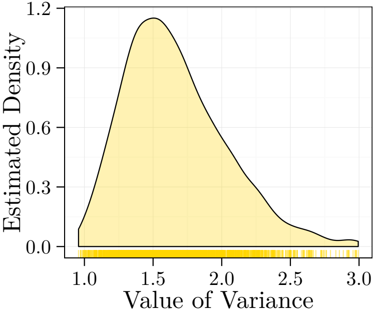
These are samples from the joint posterior, and can be used to estimate the joint posterior density.



Markov Chain Monte Carlo: Result

All the sample values for one parameter, are used to estimate the marginal density.





MCMC: Gibbs Sampler

This MCMC algorithm is called a **Gibbs sampler**.

The recipe for a bivariate sampler:

1. Find $\Pr(X|Y)$ (or something proportional).
2. Find $\Pr(Y|X)$ (or something proportional).
3. Pick starting values of (X, Y).
4. Repeat the following for desired number of samples:
   4.1 Sample new X from $\Pr(X|Y)$
   4.2 Sample new Y from $\Pr(Y|X)$
   4.3 Save these values

MCMC: Gibbs Sampler

]fragile] You will never have to code this: R has many packages that construct the Gibbs sampler for you.
The **Gibbs sampler** is clever because it takes a large multivariate problem, and turns it into several small univariate problems.

In many cases, the conditional distribution does not have a familiar form and so we use an additional method to simulate from it, ie:

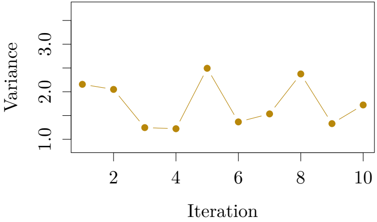
- Metropolis-Hastings step
- Slice sampler
- Data augmentation

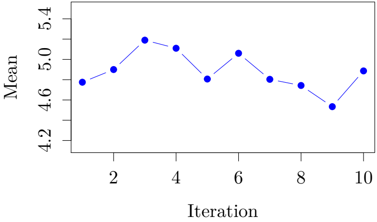
These methods are still embedded within a Gibbs sampler.

MCMC: Traceplots

A univariate view of the Gibbs sampler is the **traceplot**.

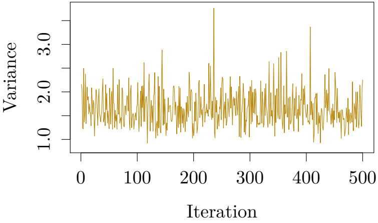
This is the typical way to view MCMC output (works for any number of parameters.) For 10 iterations:

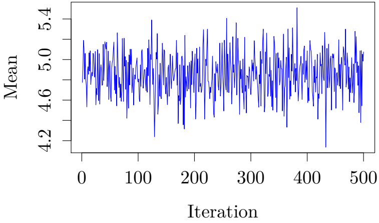




MCMC: Traceplots

For the entire Markov Chain.





Note that the chains are a time series which is **stationary** around a mean.

Ideology

Historically, there has been a divide between between **Frequentist** and **Bayesian** statistics.

Arguments about the subjectivity of priors, the validity of asymptotic assumptions

This debate is so 1990s. Now, most applied statisticians use a mix of Bayesian and Frequentist statistics.

Basically, **two reasons** why people use Bayesian statistics:

- They like the Bayesian interpretation of probability, are genuinely interested in incorporating prior information (use informative priors).
- They like MCMC and find the Bayesian framework convenient for fitting complex models (use noninformative priors).

Bayesian Modelling in R

Packages that fit pretty much anything:

- JAGS/BUGS (via R2jags/R2bugs)
- STAN (via rstan)
- MCMCpack

Packages that fit specific models:

- MCMCglmm (super cool)
- BayesLogit
- blmer

And so, so many more.