Social Data Institute

UCL

Continuing Professional Development (CPD) course

Introduction To Bayesian Inference & Modelling (June 2025)

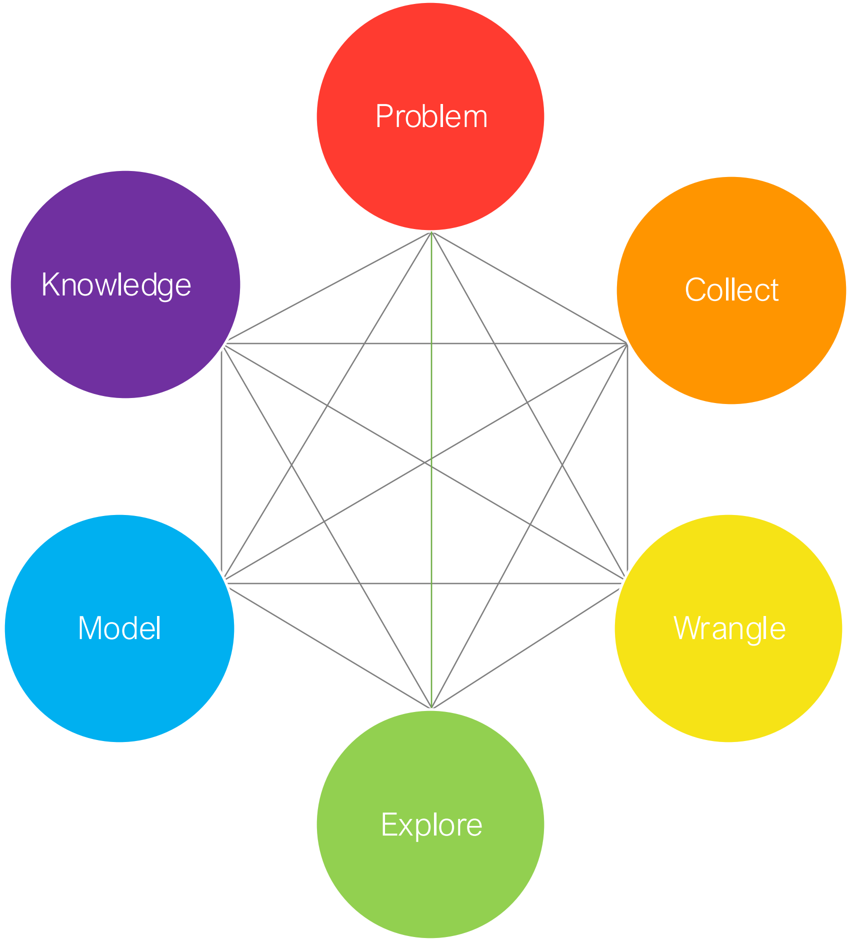
DAY 2: INTRODUCTION TO BAYESIAN INFERENCE

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Additional details:
https://www.ucl.ac.uk/social-data

Contents

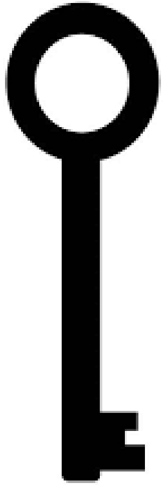
- Describe the overall workflow of Bayesian Inference
- How does the Bayes’ Rule work
- Data input, process model and assumptions
- Types of priors
- Derivation of the joint posterior distribution
- Brief examples of conjugate and non-conjugate posterior



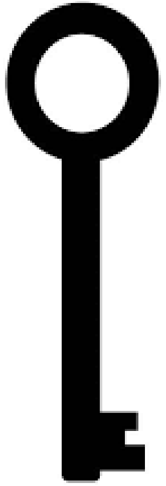
Quick recap

Definition:

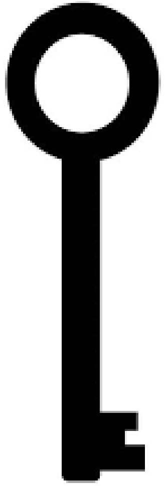
Bayesian statistics, is all about **uncertainty** (i.e., lack of complete sureness or knowledge). It is essentially the practice of expressing what you believe about something as a probability (**before observing that thing)**; and then using new evidence (**after observing that thing**) to update those beliefs thereafter.



Basic Probabilities



Probability Distribution



Bayes' Theorem

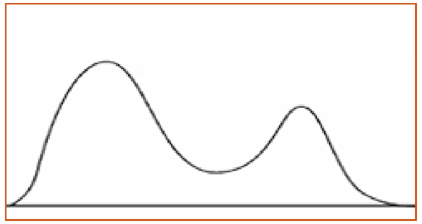
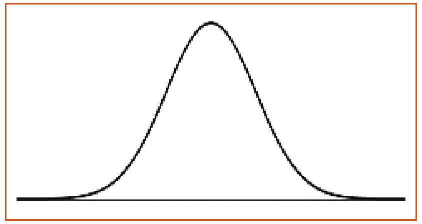
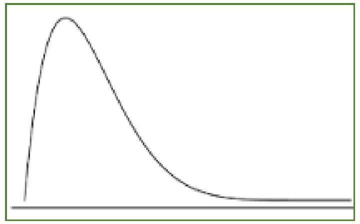
Bayesian Inference

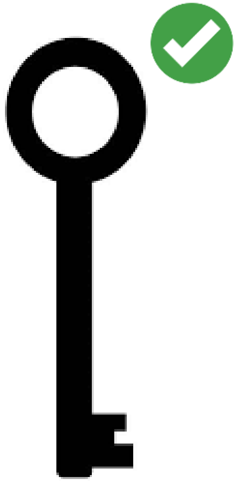
Quick recap

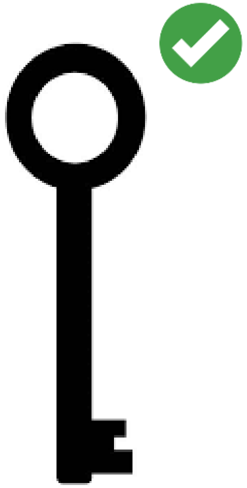
This equation is the backbone of Bayesian Statistics

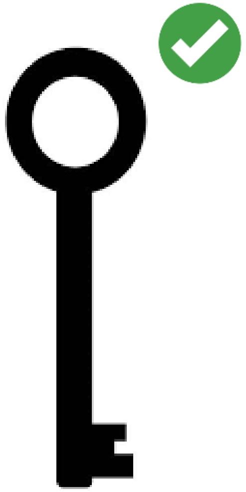
Posterior		Likelihood		Prior
Posterior: Probability of our new parameter being true **after** observing the data.		**Likelihood**: Plausibility of the observed data we have, given the parameter has a specific value(s)		**Prior**: Our belief about what the parameter is, **before** observing the data

$$\Pr(\text{Parameter} \mid \text{Data}) \propto \Pr(\text{Data} \mid \text{Parameter}) \times \Pr(\text{Parameter})$$


∝

×



Basic Probabilities


Probability Distribution


Bayes' Rule


Bayesian Inference

Bayesian Inference

Definition:

Bayesian inference uses the Bayes' Theorem to update what we believe about something (**i.e., parameters**) as we get new information (i.e., **data**). It's a way to learn and improve our understanding using both prior knowledge and new information.

Plausibility of the observed data we have, given the parameter has a specific value(s)

Our belief about what the parameter is, before observing the data

Posterior

Likelihood

Prior

Probability of our new parameter being true after we've observed dataset

$$\Pr(\text{Parameter} \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \text{Parameter})\,\Pr(\text{Parameter})}{\Pr(\text{Data})}$$

This corresponds to something out the data that's been observed

This is what is derived:

Posterior probability $\propto$ Likelihood x Prior Probability

This equation is the backbone of Bayesian Statistics

Workflow for Bayesian Inference

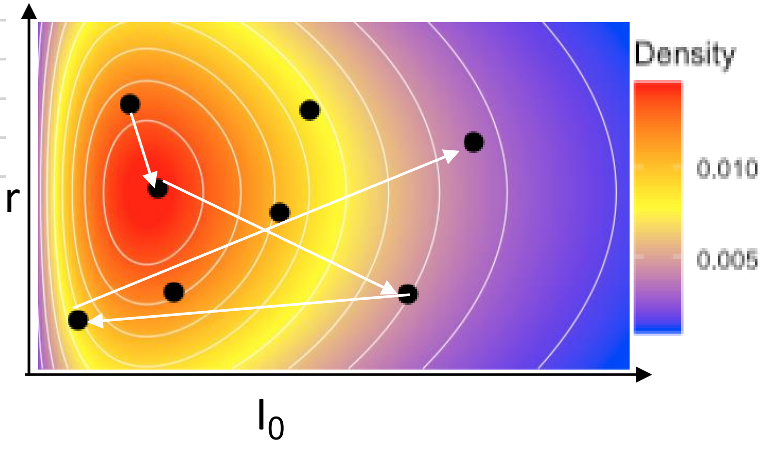
Inputs

- Observed data
- Process model for likelihood (statistical or mechanistic model)
- Building our priors i.e., assumptions about the parameters

Task 1:
Model Building

Building the Posterior Distribution

I0	r	Pr(I0, r | Data)
5	0.3	0.2
7	0.01	0.01
10	-0.02	0.15
⋮	⋮	⋮



Task 2:
Equation Sampling

Sampled Posterior

I0	r	Predictions.1
13.25199	0.2531690	22
14.56882	0.2427261	12
11.88136	0.2598835	16
10.82666	0.2669298	13
11.93687	0.2612647	14

- Sample results for parameters
- Samples for the predictions

Task 3:
Simulation the quantities of interest

Output

- Summary table (mean, percentile and credibility intervals)
- Predictions made from **generated quantities block**

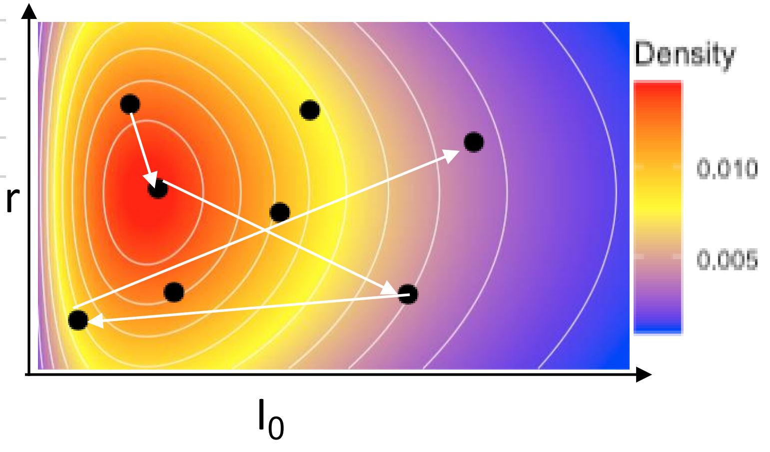
Inputs

- Observed data
- Process model for likelihood (statistical or mechanistic model)
- Building our priors i.e., assumptions about the parameters

Task 1:
Model Building

Full Posterior

I0	r	Pr(I0, r | Data)
5	0.3	0.2
7	0.01	0.01
10	-0.02	0.15
⋮	⋮	⋮



Task 2:
Equation Sampling

Sampled Posterior

I0	r	Predictions.1
13.25199	0.2531690	22
14.56882	0.2427261	12
11.88136	0.2598835	16
10.82666	0.2669298	13
11.93687	0.2612647	14

- Sample results for parameters
- Samples for the predictions

Task 3:
Simulation the quantities of interest

Output

- Summary table (mean, percentile and credibility intervals)
- Predictions made from **generated quantities block**

Inputs: Data, process model and assumptions [1]

Bayesian estimation of Epidemic Growth for Aedes-borne Infections

During the early days of heavy rainfall - there is a huge influx of mosquito populations in Recife which causes an outbreak Dengue virus. Public health officials collected daily case counts over the course of 15 days during this period where rainfalls are pronounced.

Data

t	Dengue
0	12
1	9
2	19
3	30
4	27
5	45
6	67
7	71
8	103
9	119
10	161
11	213
12	288
13	340
14	431

Process Model

$$D(t) = D_0 \times e^{r \times t}$$

Data inputs
- t (time)
- $D(t)$ i.e., Observed Dengue cases at t

Parameters
- D_0 i.e., initial number of infections
- r i.e., growth rate

Prior Assumptions

- "D_0 could be anything, but it is mostly likely 10, and probably between 1 to 20, anything higher is less likely"
- "r could be anything, it is hard to say – therefore I will be cautious and assume that it is negligible (0); however, the growth rate could be an increasing, or a decreasing value, and so I will assume a SD of 1"

Inputs: Data, process model and assumptions [2]

Broad types of priors

- ❖ **Informative Priors:**
  - ➢ These are priors that express specific and definite knowledge about the parameter we are type to estimate. These incorporate strong (and confident) beliefs that reflect existing knowledge or expert opinion.
  - ➢ These parameter ranges are assigned to higher plausibility in the spectrum in the probability's distribution spectrum

- ❖ **Weakly Informative Priors:**
  - ➢ These are priors that express some reasonable knowledge about the parameter. These incorporate modest (and not overly confident) beliefs.

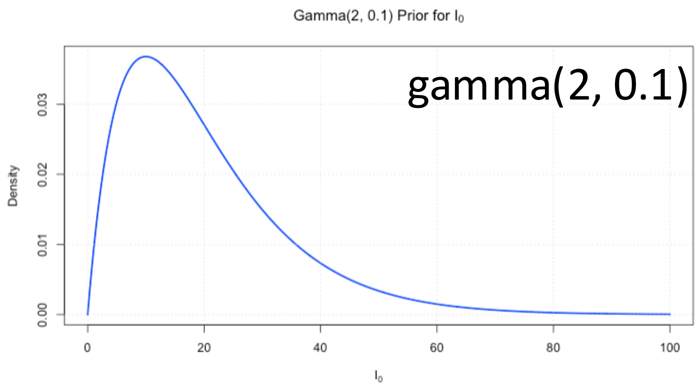
- ❖ **Uninformative (Non-informative) Priors:**
  - ➢ These are priors that express a lack of belief or knowledge about the parameter(s) in question. What often happens here are:
    - We assign uniform priors where it is assumed that all parameter values are equally likely
    - We use the default settings of the software – which decides the which priors to use.
    - Not knowing the direction of the relationship for coefficients (i.e., increasing or decreasing) or taking a neutral stance can be fall into this category of being uninformative.

Inputs: Data, process model and assumptions [3]

<table>
<tr><th>Assumptions</th><th>Probability Distribution</th><th>Prior type</th></tr>
<tr><td>

- “D_0 could be anything, but it is mostly likely 10, and probably between 1 to 20, anything higher is less likely”

</td><td>



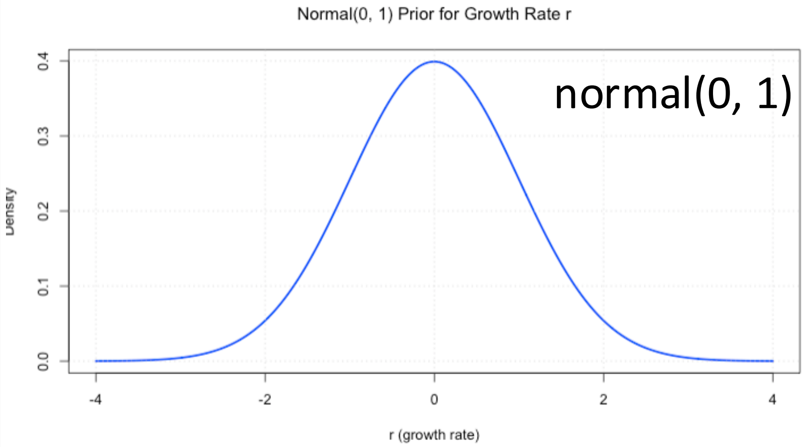
</td><td>

Weakly-Informative prior

</td></tr>
<tr><td>

- “r could be anything, it is hard to say – therefore I will be cautious and assume that it is negligible (0); however, the growth rate could be an increasing, or a decreasing value, and so I will assume a SD of 1”

</td><td>



</td><td>

Non-Informative prior

</td></tr>
</table>

Inputs

- Observed data
- Process model for likelihood (statistical or mechanistic model)
- Building our priors i.e., assumptions about the parameters

Output

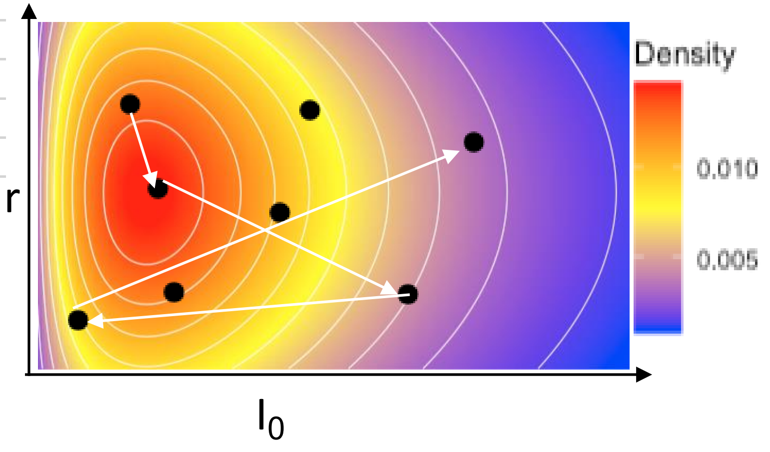
- Summary table (mean, percentile and credibility intervals)
- Predictions made from **generated quantities block**

Task 1:
Model Building

Task 3:
Simulation the quantities of interest

Full Posterior

I0	r	Pr(I0, r | Data)
5	0.3	0.2
7	0.01	0.01
10	-0.02	0.15
⋮	⋮	⋮



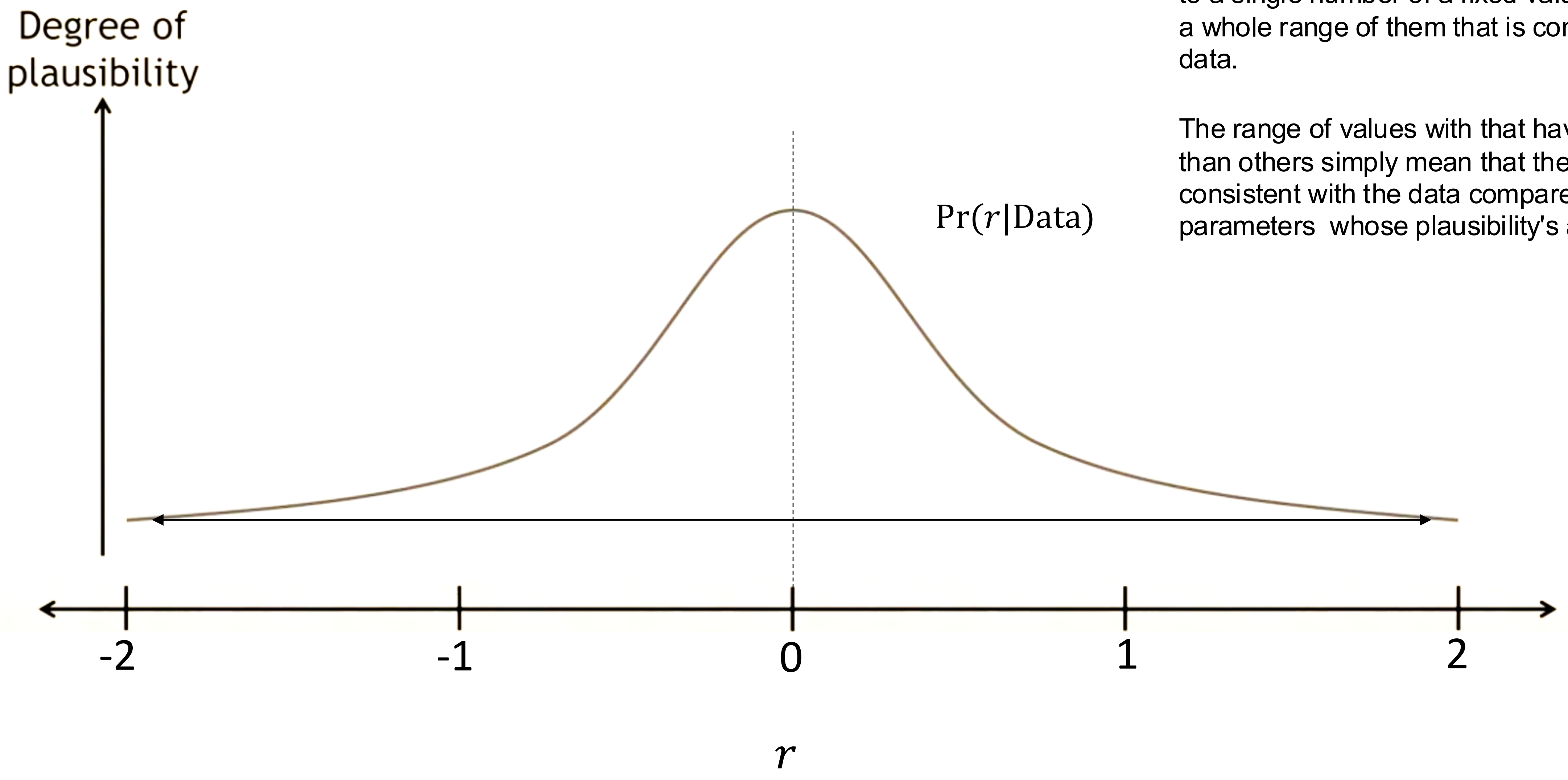
Task 2:
Equation Sampling

Sampled Posterior

I0	r	Predictions.1
13.25199	0.2531690	22
14.56882	0.2427261	12
11.88136	0.2598835	16
10.82666	0.2669298	13
11.93687	0.2612647	14

- Sample results for parameters
- Samples for the predictions

Building the Posterior Distribution [1]



Let us take epidemic growth r parameter as an example, we say we are estimating "r" parameter that is consistent with the data – we are not referring to a single number of a fixed value. We are referring a whole range of them that is consistent with the data.

The range of values with that have highly plausible than others simply mean that they're more consistent with the data compare to other parameters whose plausibility's are lower.

Building the Posterior Distribution [2]

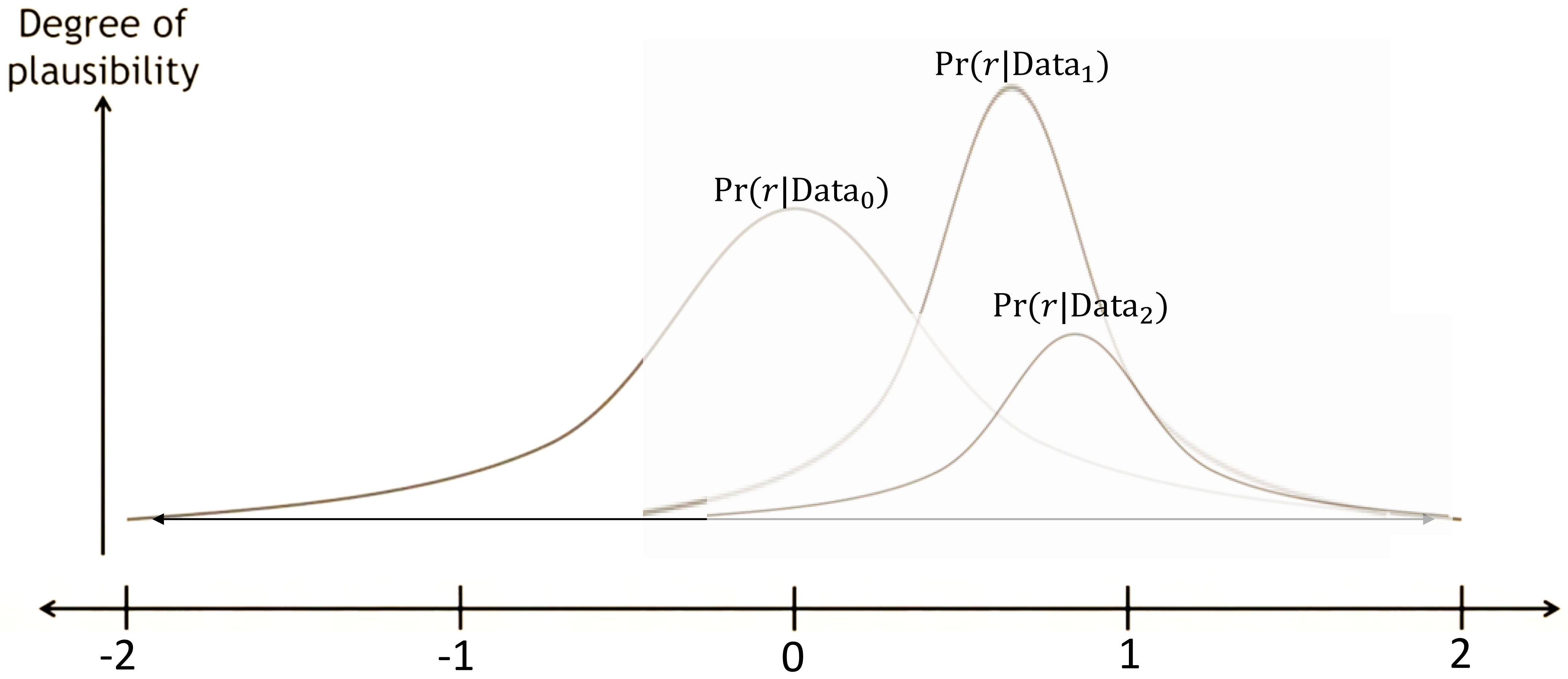
Posterior is based on a single parameter

After learning from, or conditional on this specific dataset.

$$\Pr(r|\mathrm{Data})$$

The variable (or indicator) that we are describing the plausibility about. In this case, epidemic growth rate for Dengue.

Building the Posterior Distribution [3]



Building the Posterior Distribution [4]

After learning from, or conditional on this specific dataset.

Posterior can be based on a multiple parameters.
This is called a joint posterior distribution.

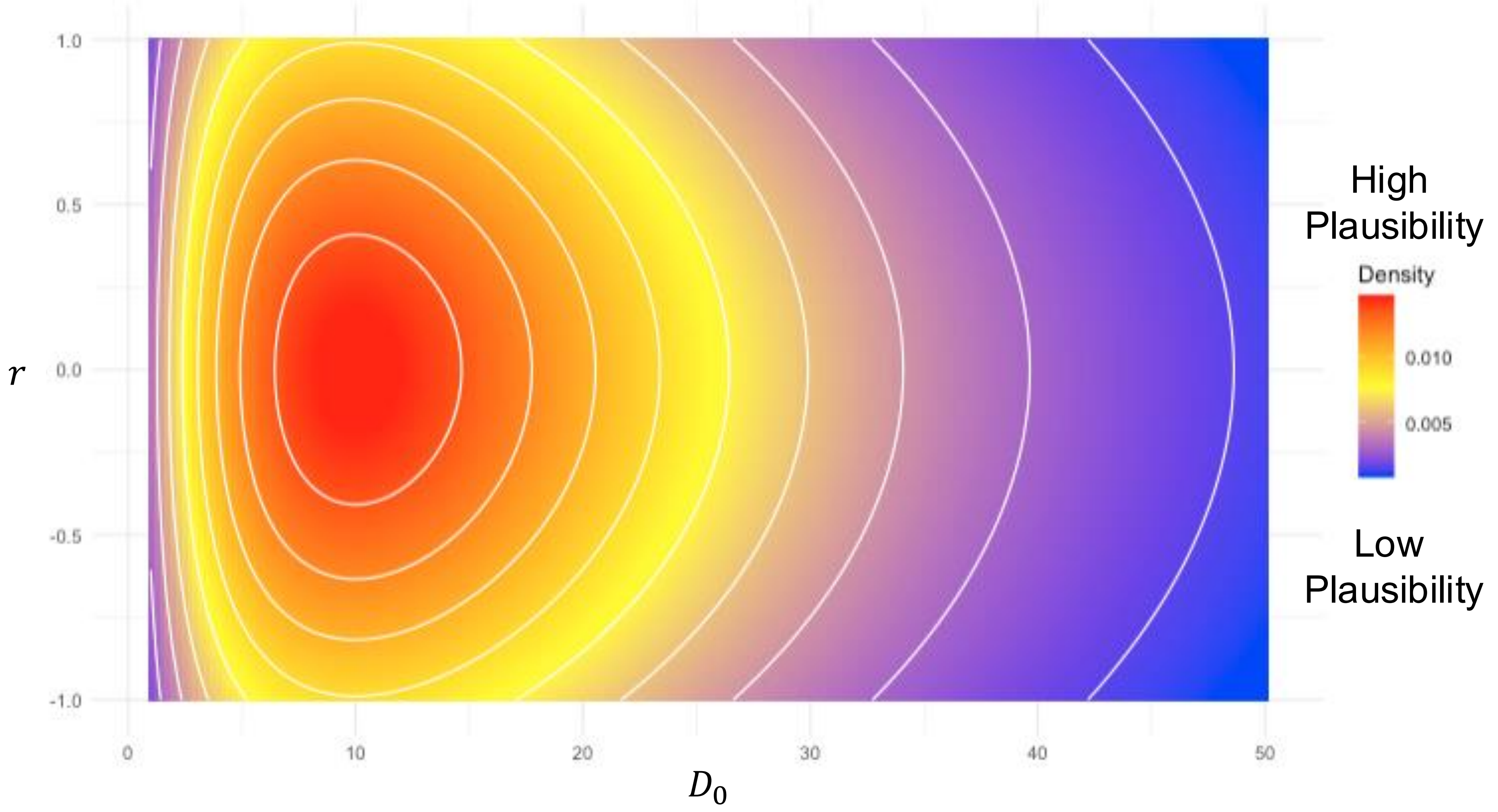
$$\Pr(D_0, r|\mathrm{Data})$$

This describes the plausibility of a combination of parameters jointly based on the data that was given. In this case, epidemic growth and initial case counts.

Building the Posterior Distribution [5]

$$\Pr(D_0, r|\text{Data})$$

- All plausibility level are between 0 and 1
- Volume under the surface is sums to 1
- Different data will change the shape of this structure

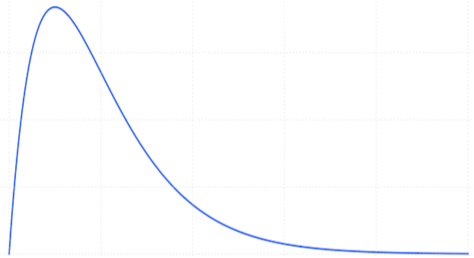


How is this posterior distribution generated?

Data

Likelihood: $\text{Poisson}(D_t \mid D_0 \times e^{r \times t})$

t	Dengue	
0	12	$\text{Poisson}(12 \mid D_0 \times e^{r \times 0})$
1	9	$\text{Poisson}(9 \mid D_0 \times e^{r \times 1})$
2	19	$\text{Poisson}(19 \mid D_0 \times e^{r \times 2})$
3	30	$\text{Poisson}(30 \mid D_0 \times e^{r \times 3})$
4	27	$\text{Poisson}(27 \mid D_0 \times e^{r \times 4})$
5	45	$\text{Poisson}(45 \mid D_0 \times e^{r \times 5})$
6	67	$\text{Poisson}(67 \mid D_0 \times e^{r \times 6})$
7	71	$\text{Poisson}(71 \mid D_0 \times e^{r \times 7})$
8	103	$\text{Poisson}(103 \mid D_0 \times e^{r \times 8})$
9	119	$\text{Poisson}(119 \mid D_0 \times e^{r \times 9})$
10	161	$\text{Poisson}(161 \mid D_0 \times e^{r \times 10})$
11	213	$\text{Poisson}(213 \mid D_0 \times e^{r \times 11})$
12	288	$\text{Poisson}(288 \mid D_0 \times e^{r \times 12})$
13	340	$\text{Poisson}(340 \mid D_0 \times e^{r \times 13})$
14	431	$\text{Poisson}(431 \mid D_0 \times e^{r \times 14})$



$D_0 \sim \text{gamma}(2, 0.1)$

$\Pr(D_0)$

$r \sim \text{normal}(0, 1)$

$Pr(r)$

Prior distribution built from set of assumption

The process model is what we build as the likelihood function. For each data point we observed, we find its likelihood function and multiply them all together

Posterior ∝

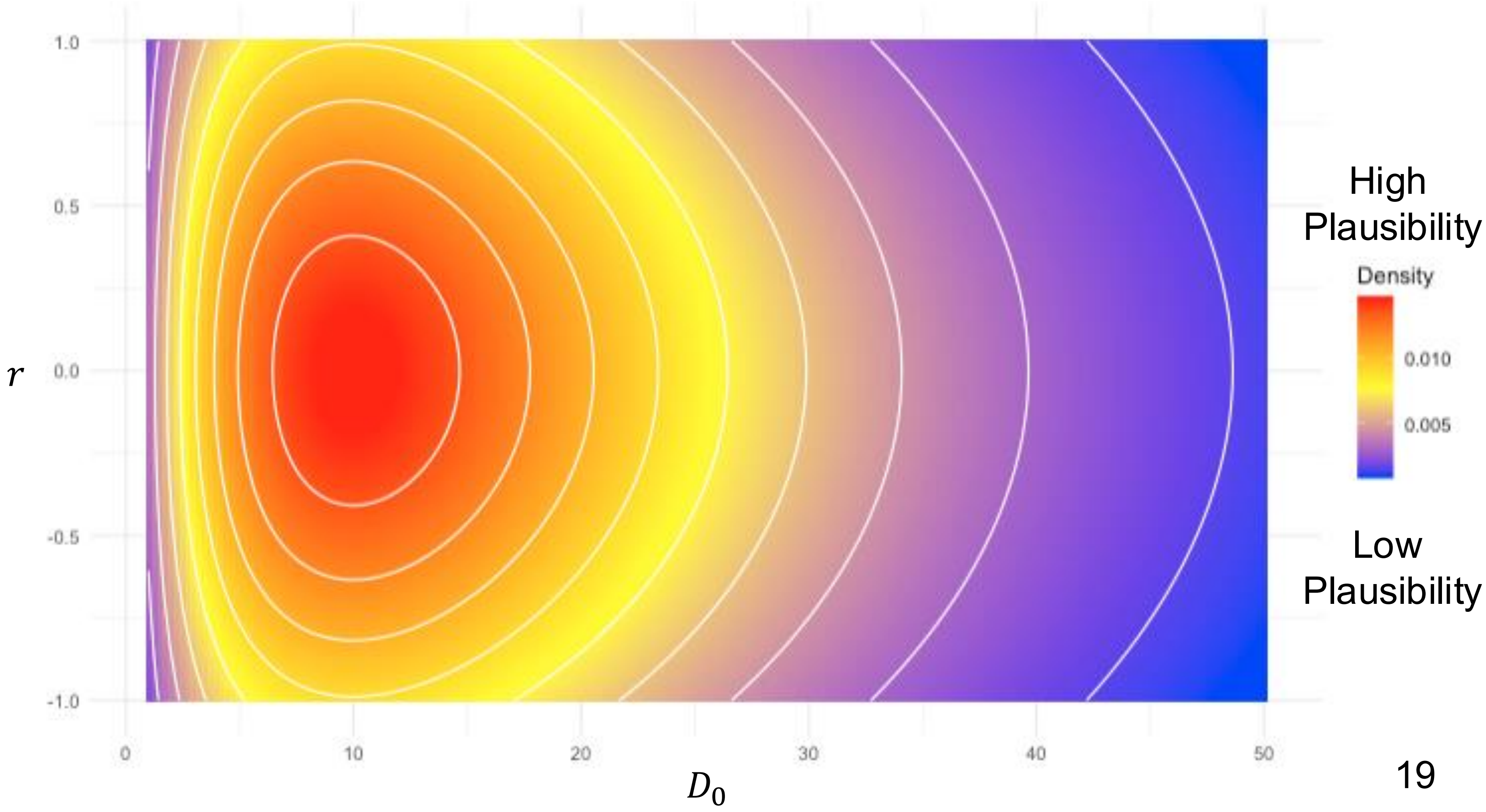
Likelihood function

$\text{Poisson}(12 \mid D_0 \times e^{r\times 0})\text{Poisson}(9 \mid D_0 \times e^{r\times 1}) \times \cdots \times \text{Poisson}(431 \mid D_0 \times e^{r\times 14}) \times$

Priors

$\text{gamma}(D_0; 2, 0.1) \times \text{normal}(r; 0, 1)$

$\Pr(D_0, r \mid \text{Data})$



This example is a **non-conjugate** posterior. This is a result formed by combining different types of probability distributions together to form a custom distribution.

Motivating example [1]:

$$\Pr(\theta \mid y) = \frac{\Pr(\theta)\ \Pr(y \mid \theta)}{\Pr(y)} \propto \Pr(\theta)\ \Pr(y \mid \theta)$$

- What is the probability (or prevalence) of infestation in Recife this year?

Survey year: 05/2023	Number properties detected with Aedes	Overall number of properties surveyed
* Most recent data collection effort.	428	976

Important Information:

- y represent the number of infested properties (428)
- n represent the overall number of properties surveyed (976)
- θ represent the unknown probability (or prevalence) of infestation
- Prior information for θ (i.e., our knowledge or belief) is assumed 0.20 (in most cases, the prevalence from past research is often this value of 20-25%.

Likelihood function: $P(y \mid \theta) = \binom{n}{y}\ \theta^{y}(1-\theta)^{n-y}$

Alternatively, instead type this mathematical notation, you use this statistical notation: $y \sim \mathrm{Bin}(n, \theta)$

- We have specified the probability function for the likelihood, what about the function for the prior i.e., $P(\theta)$?

dbeta(prop, $\alpha - 1$, $\beta - 1$)
pbeta(prop, $\alpha - 1$, $\beta - 1$)

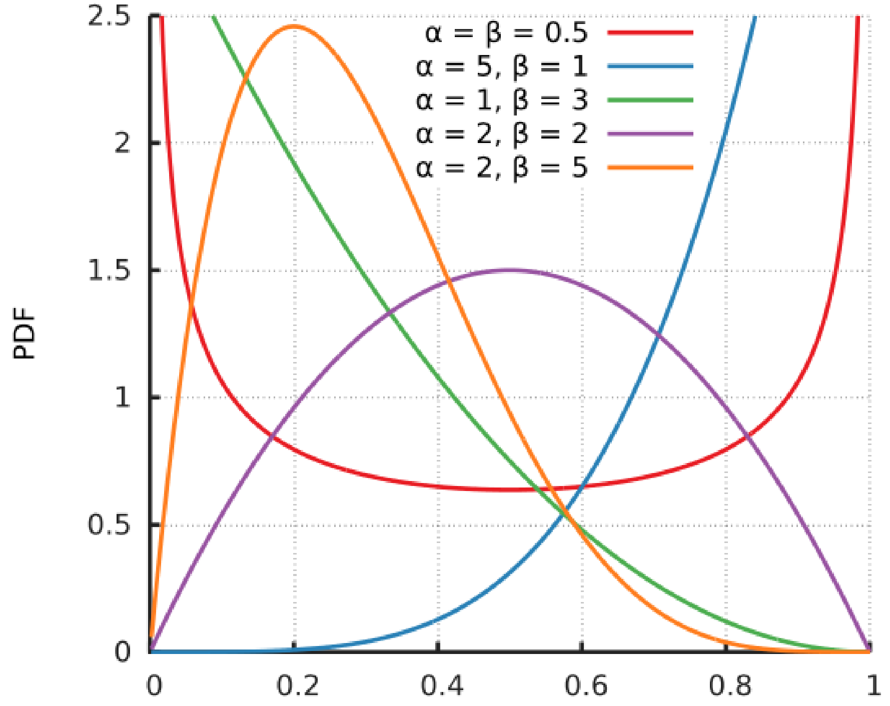
Motivating example [2]:

- The probability distribution needed in this situation is a Beta distribution. It is the best probability function to use a prior distribution for unknown parameter that's a proportion.

$$P(\theta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, (0 < \theta < 1)$$

$$P(\theta) = \text{beta}(\theta|\alpha - 1, \beta\ - 1)$$

```
Posterior mean:           qbeta(0.5,   α − 1,  β − 1)
Posterior lower limit:    qbeta(0.025, α − 1,  β − 1)
Posterior upper limit:    qbeta(0.975, α − 1,  β − 1)
```



- Example of a flexible PDF function as we can bend it to accordingly be setting values to α and β
- Here, we need to use values for α and β which gives us a distribution with a shape that's concentrated on 20-25%.
- This type of prior is an **informative prior**, because we've assigned a distribution with information that's specific.

Solutions [1]:

- Likelihood function: $P(y \mid \theta) = \binom{n}{y} \theta^{y}(1-\theta)^{n-y}$
- Prior: $P(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

- For a binomial model specification, with Beta prior, the posterior is:

$$
\begin{aligned}
P(\theta \mid y) \propto\ & P(\theta)\, P(y \mid \theta) \\
& \propto [\theta^{\alpha-1}(1-\theta)^{\beta-1}][\theta^{y}(1-\theta)^{n-y}] \\
& \Rightarrow \theta^{\alpha-1}\theta^{y}(1-\theta)^{\beta-1}(1-\theta)^{n-y} \\
& \Rightarrow \theta^{(\alpha+y)-1}\,(1-\theta)^{(\beta+n-y)-1} \\
= \mathrm{Beta}(\theta|\ & (\alpha+y)-1, (\beta+n-y)-1)
\end{aligned}
$$

All constants with the parameter of interest can be removed i.e., $\binom{n}{y}$ and $\frac{1}{B(\alpha,\beta)}$

Arranging like terms so θ's and $(1-\theta)$'s together

Using indices $a^m a^n = a^{m+n}$

- As you can see the posterior distribution is a Beta distribution, but the parameters for it have been updated.
- When the posterior is in the same family as the prior, we say the prior is a conjugate for the model.
- Here, the Beta prior is a conjugate for the binomial model.

Solutions [2]:

Updated Bayesian model: $P(\theta|Y) \propto \theta^{(\alpha+y)-1}(1-\theta)^{(\beta+n-y)-1}$

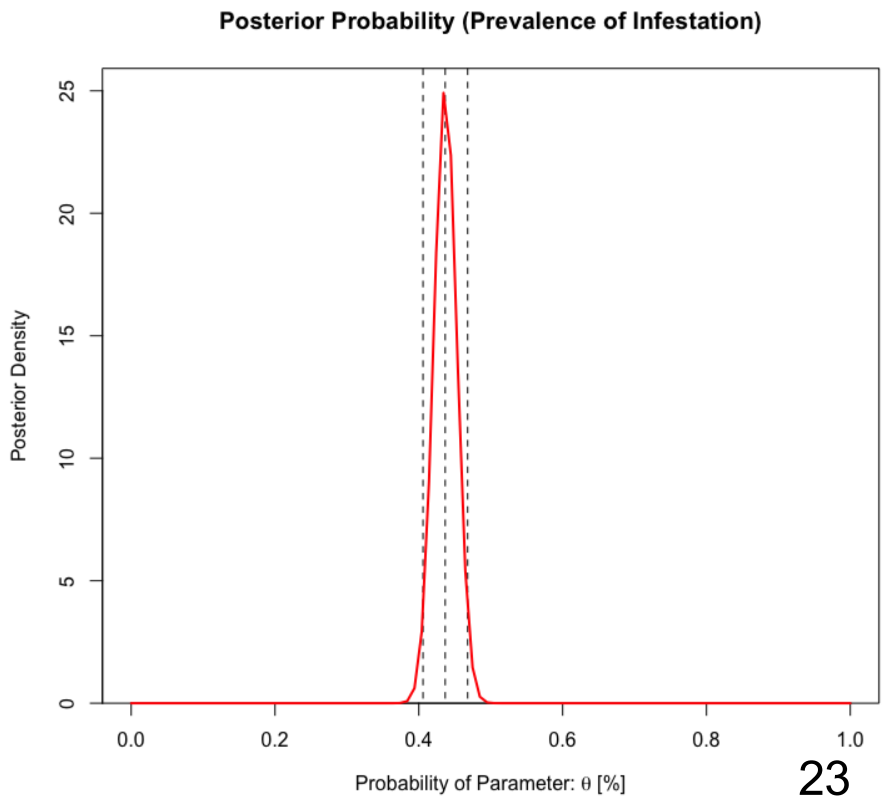
- Combining this with infestation data, the posterior is $\mathrm{Beta}(\theta|\ (\alpha+y)-1, (\beta+n-y)-1)$
- Using an informative prior in this situation, where α = 1 and β = 5, we get the following

$$\mathrm{Beta}(\theta|\ (1+428)-1, (5+976-428)-1)$$

- Above model will generate a posterior distribution; where, θ is from 0 to 1.

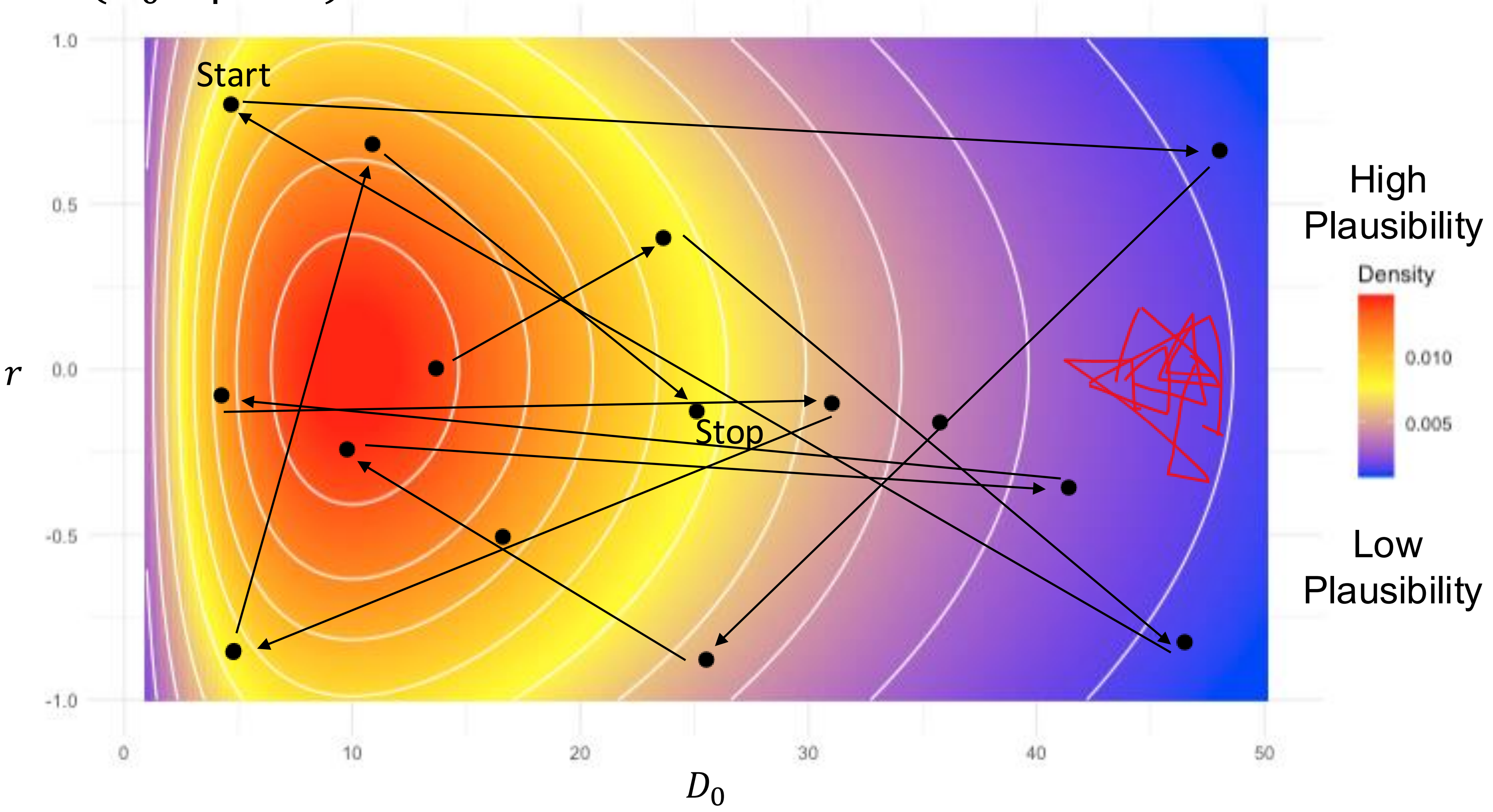
Posterior mean: qbeta(0.5, $(1+428-1,\ (5+976-428)\ -1)$
Posterior lower limit: qbeta(0.025, $(1+428-1,\ (5+976-428)\ -1)$
Posterior upper limit: qbeta(0.975, $(1+428-1,\ (5+976-428)\ -1)$

- The mean prevalence is approximately 44% with 95% credible intervals (40% to 47%)



Sampling the quantities of interest [1]

$\Pr(D_0, r|\text{Data})$



We use various algorithms to generate a posterior sample from the full joint distribution so that its manageable.

Markov Chains Monte Carlo
No U-turn
Hamiltonian Monte Carlo

The patterns of sampling should be like this illustration shown with the black lines. Where it is moving everywhere in an even manner. Otherwise, if the sampling is cluster in a particular region like the red lines – then it is biased leading divergent samples.

Any questions?