Agenda

- Bayesian analysis in a nutshell
- Data-Generating Process (DGP) and Bayes's rule
- A simple analysis of count data
- An example using the `rstanarm` package

Bayesian analysis in a nutshell

a way of statistical modelling that treats all unknowns as random variables

Bayesian analysis in a nutshell

"unknowns" can include:

- parameters of the distributions
- regression coefficients
- missing values in the data
- future observations
- ...

Bayesian analysis in a nutshell

treating all unknowns as random variables = assigning probabilities to all (subsets of) values they can take

Bayesian analysis in a nutshell

A model of data-generation with unknown parameters
+ prior (initial) beliefs about parameters
+ data

posterior (updated) beliefs about parameters

Predictions

Summaries & quantities of interest

Inputs for further analyses

Why Bayesian?

Bayesian analysis will help you if

- you wish to interpret the intervals around point estimates in terms of probabilities of different values
  - credible intervals vs confidence intervals
- you wish to include information from other studies
  - or do meta-analysis
- you wish to plug the estimates into cost-benefit analyses
- you want to estimate a somewhat complicated model
- the number of observations is too small for conventional hypotheses testing

Data-generating process

- A chain of (hypothetical) events that produced each observation in the data
- ...can include both deterministic and random nodes
- ...can include unknown parameters

** an abstraction! **

Logit: DGP

$$\mathbf{x}_i \longrightarrow \blacksquare \longrightarrow y_i = 0 \text{ or } y_i = 1$$

Logit: DGP

1. coefficients $\boldsymbol{\beta}$ and the values of independent variables $\mathbf{x}_i$ determine θ_i

$$\theta_i = \text{logit}^{-1}\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}\right)$$

2. Nature returns $y_i = 1$ with probability (θ_i) and $y_i = 0$ with probability $(1 - \theta_i)$

- we observe $\mathbf{x}_i$, y_i
- and wish to learn about $\boldsymbol{\beta}$

Logit: Likelihood function

1. coefficients $\boldsymbol{\beta}$ and the values of independent variables $\mathbf{x}_i$ determine θ_i

$$\theta_i = \text{logit}^{-1}\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}\right)$$

2. Nature returns $y_i = 1$ with probability (θ_i) and $y_i = 0$ with probability $(1 - \theta_i)$

Likelihood function: $L(\boldsymbol{\beta}|\text{data}) = p(\text{data}, \boldsymbol{\beta})$

$x_{i,1}$	$x_{i,2}$	y_i	$p(y_i\|\boldsymbol{\beta}, \mathbf{x}_i)$
1	-1	1	$\text{logit}^{-1}(\beta_0 + \beta_1 - \beta_2)$
2	1	0	$1 - \text{logit}^{-1}(\beta_0 + 2\beta_1 + \beta_2)$
0	3	1	$\text{logit}^{-1}(\beta_0 + 3\beta_2)$

$$L(\boldsymbol{\beta}|\text{data}) = \text{logit}^{-1}(\beta_0 + \beta_1 - \beta_2) \cdot \left(1 - \text{logit}^{-1}(\beta_0 + 2\beta_1 + \beta_2)\right) \cdot \text{logit}^{-1}(\beta_0 + 3\beta_2)$$

A non-Bayesian approach

1. Assume that the true values of parameters β exist
2. Formulate likelihood function with the data at hand
3. Find $\hat{\beta}$ that maximizes this function
4. Treat $\hat{\beta}$ as the sample estimate of the true β

Bayesian approach

1. Treat parameters $\boldsymbol{\beta}$ as random variables
   1. Figure out which values they can take
   2. Assign a distribution to its values (prior distribution)
2. Use data to update the beliefs about the distribution of $\boldsymbol{\beta}$ (this new distribution is called posterior)
   1. Formulate likelihood function with the data at hand
   2. Apply Bayes's rule:

$$\underbrace{p(\boldsymbol{\beta}|\text{data})}_{\text{posterior}} \propto \overbrace{p(\text{data}|\boldsymbol{\beta})}^{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\beta})}_{\text{prior}}$$

Learning about weather by looking into window

Parameter:
weather $\theta \in \{\text{cold}, \text{warm}, \text{hot}\}$
DGP:
If it is cold, a person wears a coat with probability 0.7 and wears sandals with probability 0.3.
If it is warm, a person wears a coat with probability 0.5 and wears sandals with probability 0.5.
If it is hot outside, a person sandals with probability 0.2 and wears sandals with probability 0.8.
Data:
4 in coats, 1 in sandals

Likelihood calculations

Parameter:

weather $\theta \in \{\text{cold}, \text{hot}\}$

DGP:

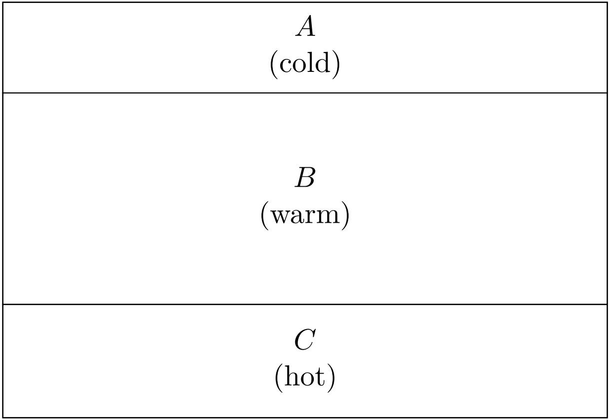
$$p(\text{coat} \mid \text{cold}) = 0.7 \text{ and } p(\text{sandals} \mid \text{cold}) = 0.3$$
$$p(\text{coat} \mid \text{warm}) = 0.5 \text{ and } p(\text{sandals} \mid \text{warm}) = 0.5$$
$$p(\text{coat} \mid \text{hot}) = 0.2 \text{ and } p(\text{sandals} \mid \text{hot}) = 0.8$$

Data:

4 in coats, 1 in sandals

$$\begin{aligned}
L(\text{cold} \mid \text{data}) &= c(5,4) \cdot p(\text{coat} \mid \text{cold})^4 \cdot p(\text{sandals} \mid \text{cold})^1 \\
&= c(5,4) \cdot 0.7^4 \cdot 0.3^1 \approx 0.360 \\
L(\text{warm} \mid \text{data}) &= c(5,4) \cdot p(\text{coat} \mid \text{warm})^4 \cdot p(\text{sandals} \mid \text{warm})^1 \\
&= c(5,4) \cdot 0.5^4 \cdot 0.5^1 \approx 0.156 \\
L(\text{hot} \mid \text{data}) &= c(5,4) \cdot (\text{coat} \mid \text{hot})^4 \cdot p(\text{sandals} \mid \text{hot})^1 \\
&= c(5,4) \cdot 0.2^4 \cdot 0.8^1 \approx 0.006
\end{aligned}$$

Prior



$$p(\text{cold}) = \frac{A}{A+B+C}$$

Prior × Likelihood

$A \cdot p(\text{sandals} \mid \text{cold})$

cold ∩ coat
$A \cdot p(\text{coat} \mid \text{cold})$

cold ∩ sandals

warm ∩ coat
$B \cdot p(\text{coat} \mid \text{warm})$

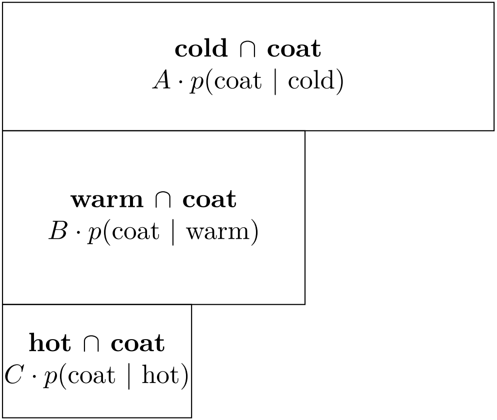
warm ∩ sandals
$B \cdot p(\text{sandals} \mid \text{warm})$

hot ∩ coat
$C \cdot p(\text{coat} \mid \text{hot})$

hot ∩ sandals
$C \cdot p(\text{sandals} \mid \text{hot})$

Posterior

Observed 1 person in coat:



$$p(\text{cold} \mid \text{coat}) = \frac{A\, p(\text{coat} \mid \text{cold})}{A\, p(\text{coat} \mid \text{cold}) + B\, p(\text{coat} \mid \text{warm}) + C\, p(\text{coat} \mid \text{hot})}$$

$$p(\text{cold} \mid \text{data}) \propto p(\text{cold})\, p(\text{data} \mid \text{cold})$$
$$p(\text{warm} \mid \text{data}) \propto p(\text{warm})\, p(\text{data} \mid \text{warm})$$
$$p(\text{hot} \mid \text{data}) \propto p(\text{hot})\, p(\text{data} \mid \text{hot})$$

With the DGP and data as before:

$$p(\text{data} \mid \text{cold}) \approx 0.360$$
$$p(\text{data} \mid \text{warm}) \approx 0.156$$
$$p(\text{data} \mid \text{hot}) \approx 0.006$$

and non-informative priors:

	cold	warm	hot
Prior	0.333	0.334	0.333
Likelihood	0.360	0.156	0.006
Posterior	0.689	0.299	0.011

$$p(\text{cold} \mid \text{data}) \propto p(\text{cold})\, p(\text{data} \mid \text{cold})$$
$$p(\text{warm} \mid \text{data}) \propto p(\text{warm})\, p(\text{data} \mid \text{warm})$$
$$p(\text{hot} \mid \text{data}) \propto p(\text{hot})\, p(\text{data} \mid \text{hot})$$

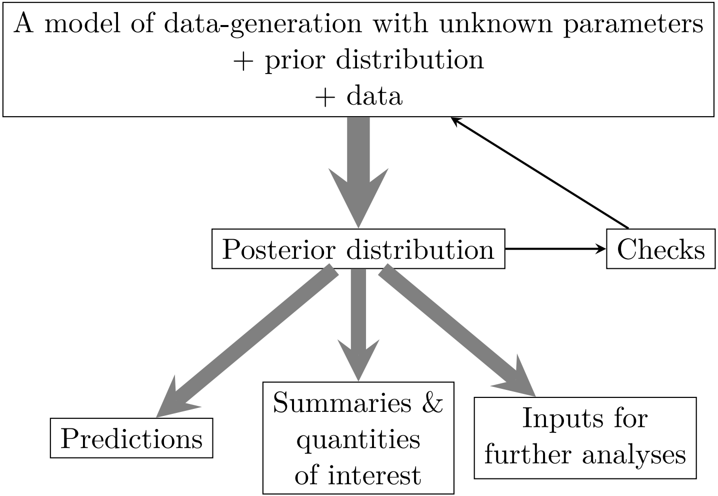
With the DGP and data as before:

$$p(\text{data} \mid \text{cold}) \approx 0.360$$
$$p(\text{data} \mid \text{warm}) \approx 0.156$$
$$p(\text{data} \mid \text{hot}) \approx 0.006$$

and more informative priors:

	cold	warm	hot
Prior	0.01	0.09	0.9
Likelihood	0.360	0.156	0.006
Posterior	0.156	0.609	0.234

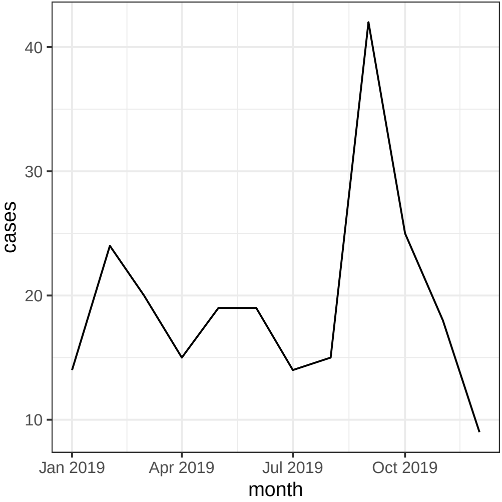
Workflow


A model of data-generation with unknown parameters
+ prior distribution
+ data
Posterior distribution
Checks
Predictions
Summaries & quantities of interest
Inputs for further analyses

Number of bike thefts in Exeter

Data:

month	cases
Jan-19	14
Feb-19	24
Mar-19	20
Apr-19	15
May-19	19
Jun-19	19
Jul-19	14
Aug-19	15
Sep-19	42
Oct-19	25
Nov-19	18
Dec-19	9

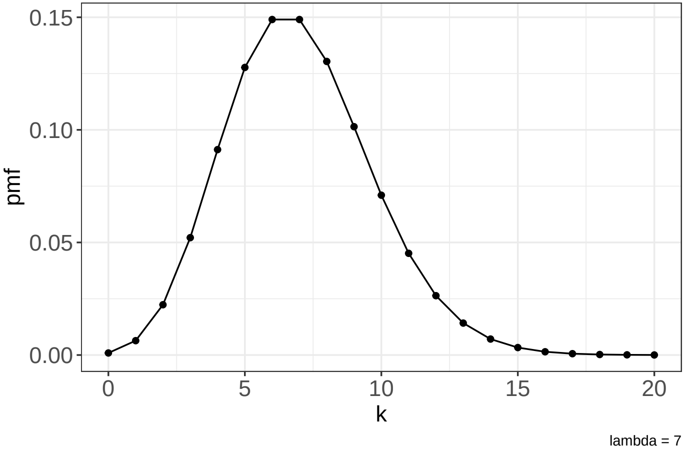


DGP

Each month's value of k_i is randomly drawn from the Poisson distribution with the location parameter at λ

Poisson distribution:

$$p(k) \propto \lambda^k e^{-\lambda}$$



Likelihood function

DGP: Each month's value of k_i is randomly drawn from the Poisson distribution with the location parameter at λ

$$p(k_i|\lambda) \propto \lambda^{k_i} e^{-\lambda}$$

$$p(\text{data}|\lambda) \propto \prod_{i=1}^{N} \lambda^{k_i} e^{-\lambda} = \lambda^{\sum_{i=1}^{N} k_i} e^{-N\lambda}$$
$$\propto \lambda^{234} e^{-12\lambda}$$

Priors

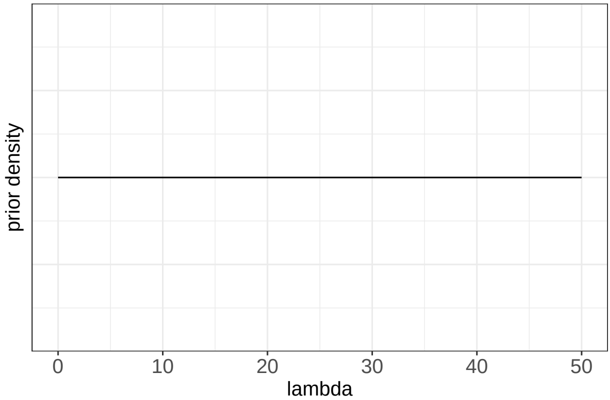
An opportunity to

- incorporate qualitative information or prior knowledge into estimation
- restrict the range for the parameter search
- reduce the computational resources needed for estimation

Prior 1

Uniform on $(0, \infty)$: $p(\lambda) \propto 1$ if $\lambda > 0$

- non-informative
- improper (does not integrate to 1)
- $p(\lambda \leq 0) = 0$



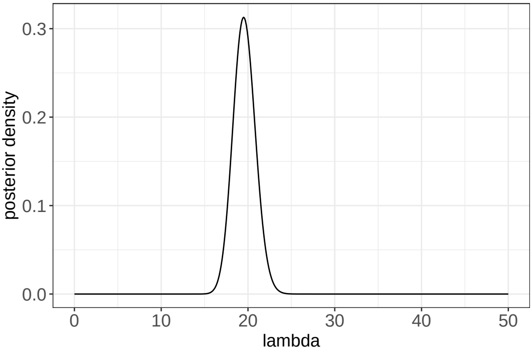
Posterior 1

Prior: $p(\lambda) \propto 1$

Likelihood: $p(\text{data}|\lambda) \propto \lambda^{234} e^{-12\lambda}$

Posterior: $p(\lambda|\text{data}) \propto \lambda^{234} e^{-12\lambda} \cdot 1 = \lambda^{234} e^{-12\lambda}$

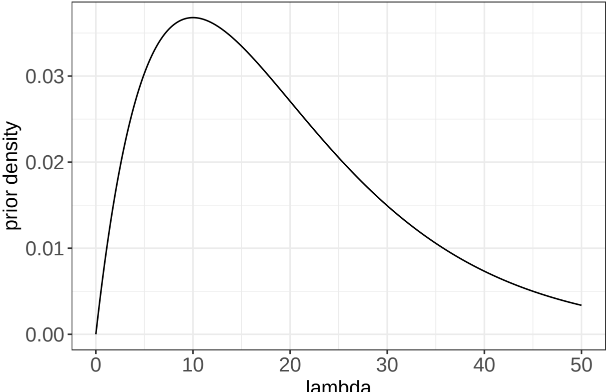
The posterior distribution of λ is a gamma distribution with shape=235 and rate=12



Prior 2

Gamma distribution with shape=2 and rate=0.1:
$p(\lambda) \propto \lambda e^{-0.1\lambda}$

- weakly informative
- Gamma prior is conjugate to Poisson likelihood – makes it easy to compute the posterior
- $p(\lambda < 0) = 0$



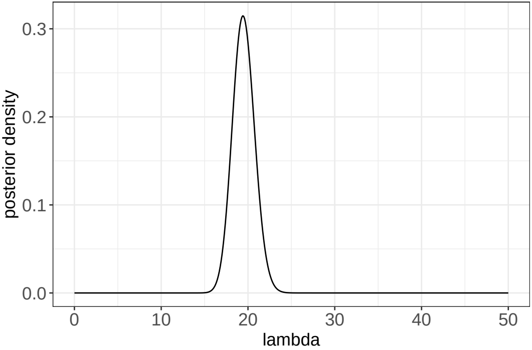
Posterior 2

Prior: $p(\lambda) \propto e^{-0.1\lambda}$

Likelihood: $p(\text{data}|\lambda) \propto \lambda^{234} e^{-12\lambda}$

Posterior: $p(\lambda|\text{data}) \propto \lambda^{234} e^{-12\lambda} \cdot \lambda e^{-0.1\lambda} = \lambda^{235} e^{-12.1\lambda}$

The posterior distribution of λ is a gamma distribution with shape=236 and rate=12.1



Another way of recording the posterior distribution

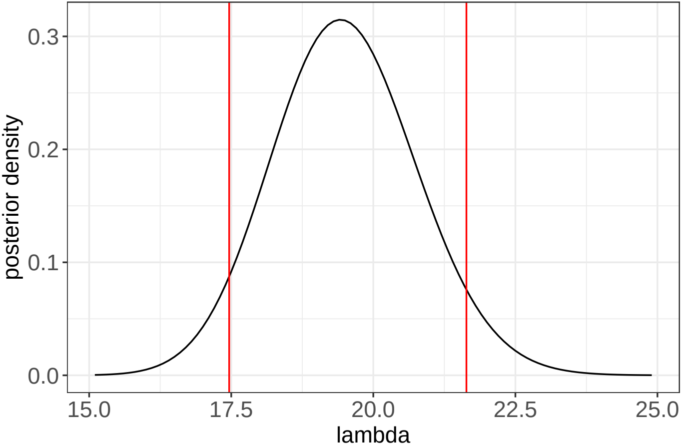
	lambda
1	16.95578167
2	19.22808885
3	19.13343529
4	18.89169685
5	18.70425733
6	18.96495068
7	17.50372344
8	21.25867386
9	19.25259324
10	19.36690632
11	19.46680601
12	20.60875773
13	20.89049706
14	18.37696661
15	19.26760371
16	17.7946507
17	18.86570533
...	...

Number of columns = number of parameters

Number of rows = number of random draws from posterior distribution

What to do with the posterior distribution?

Plot it:



What to do with the posterior distribution?

Summarize it:

- Point estimates
  - Posterior mean
  - Posterior mode
- Credible intervals
  - Central intervals (a 90% credible interval spans between 5th and 95th percentile)
  - Highest posterior density intervals

* in the bike theft example:
Mean ($=\frac{\text{shape}}{\text{rate}}$): 19.58
Mode ($=\frac{\text{shape}+1}{\text{rate}}$): 19.67
90% credible interval: between 17.46 and 21.64

What to do with the posterior distribution?

Compute posterior distributions for the quantity of interests

- Conditional probabilities in logit/probit models
- Marginal effects
- ...

What to do with the posterior distribution?

Compute predictive distributions for the "data" that does not exist:

Simulations method:

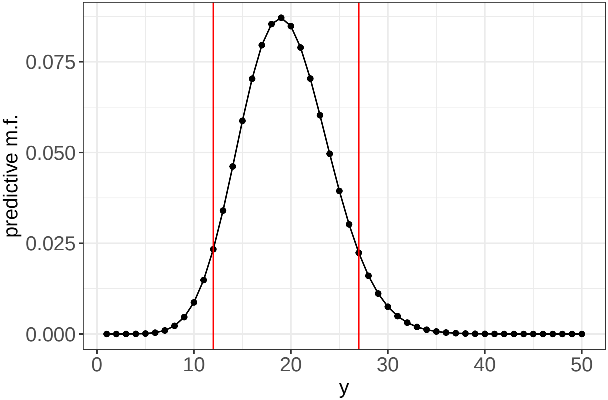
- sample parameters from the posterior distribution
- apply DGP to produce a new observation
- repeat and summarize

Analytical method

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}|\theta)\, p(\theta|y) d\theta$$

Prediction

In the bike thefts example



Mean: 19.5

Mode: 19

90% predictive interval is from 12 to 27

Another way of computing the posterior distribution

MCMC (Markov chain Monte Carlo) / Metropolis algorithm:

1. Pick a point in the parameter space
2. Sample a proposed move (to another point) from the proposal distribution
3. Use the ratio of posterior densities to accept/reject the proposed move
4. Repeat 2 and 3 until the distribution of the visited points in the parameter space does not depend on the path

The distribution of the visited points in the parameter space converges to the target posterior distribution

Stand-alone programmes for running MCMC:

- BUGS
- WinBUGS
- JAGS
- OpenBUGS
- Stan

(typically fed with data and model from R or Python)

R Packages for specifying Bayesian models command-style:

- brms
- rstanarm
- MCMCpack

rstanarm package

- relies on the Stan implementation of HMC for estimation
- includes commonly used "frequentist" models
  - the syntax of stan_lm() mimics the syntax of lm()
  - the syntax of stan_glm() mimics the syntax of glm()
  - the syntax of stan_glmer() mimics the syntax of glmer()
- uses weakly informative priors as defaults
- produces MCMC draws and summaries