Introduction
to
Bayesian Analysis

Learning Objectives

- You will understand the principles underlying basic Bayesian modelling.

- You will be able to differentiate the principles of both classical and Bayesian statistical frameworks, and advantages / disadvantages of each.

- You will learn how to use Bayesian analysis for making inference about real-world problems.

- Use software such as R, JAGS (STAN) to implement Bayesian analyses.

(More details on computational techniques for Bayesian analysis and forming hierarchical statistical models will be dealt in Bayes II course.)

Course Information

- Textbook:

Bayesian data analysis by A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, Chapman & Hall.

- References

Bayesian Statistical Modelling, Peter Congdon

Bayesian Statistics: an introduction, Peter M. Lee

- You are responsible for reading about materials related to lecture contents in some book: google …

- Statistics workshop/Työpaja : Wed 16:15-17:45 by Mikko Korhonen

Grade Information

- Evaluation: Exercises (20%) R/JAGS- assignments (10%) Final Exam (70%)

- **Lab assignments** : due within a week after each lab session

- Will check only how many exercise problems you have tried, **not** whether your answers are correct or not. You can collaborate with others.

- Extra credit from Lab participation (on-site) or Early submission

- **R & JAGS assignments**: must be done individually - No collaboration allowed
  You can ask the instructor for guidance. (Python/Stan can be used)

- **Final exam**: 18.12 (13-16) You need to register for the exam in SISU.

If you fail, you can do either retake (written) exam for a max. grade 5, or

take the online computer-analysis exam for a max. grade 3.

Lecture plan (subject to change)

21.10.24	Introduction: Bayes' Theorem, Frequentists vs. Bayesian school	
23.10.24	Likelihood, kernel of a density, Normalizing constants	Lab 0
28.10.24	Conjugate priors and updates	
30.10.24	Choice of priors, posterior inference	Lab 1
04.11.24	Normal samples (one parameter inference)	
06.11.24	Posterior summaries (R examples)	Lab 2
11.11.24	Predictive distributions/inference	
13.11.24	Mixture of conjugate priors	Lab 3 (R)
18.11.24	Normal sample with two unknown parameters	
20.11.24	Introduction to BUGS (RJAGS) **	Lab 4
25.11.24	Two Normal samples/Linear Models	
27.11.24	Multiparameter distributions	Lab 5(BUGS)
02.12.24	Hierarchical models, Empirical Bayes	
04.12.24	Review	Lab 6

Remarks before we begin

- In the beginning, the difficult part of Bayesian inference is learning to think in a Bayesian manner.

- People who have already learned some classical statistics may experience

some difficulty in shifting to a Bayesian mode of thinking.

Some things may have to be “unlearned”.

- However, once you “get it,” you’ll find it much easier to understand standard statistical ideas as well as the Bayesian ones.

So, it is an advantage for a classical statistician to understand Bayesian statistics.

Probability
and
Bayes theorem

Bayesian statistics

- `Bayesian': named after Thomas Bayes (1702-1761)

- The basic tool of inference is **Bayes' Theorem** (Bayes' Law or Bayes' Rule).

Bayesian statistics combines prior information and sample data to make conclusions about a parameter of interest.

- Bayesian inference differs from classical **frequentist inference** in that it specifies a probability distribution for the parameter(s) of interest.

- Bayesian statistical methods are widely used in many science and engineering areas:

decision theory, estimation and prediction, machine learning & AI, expert systems, medical imaging, pattern recognition, data compression and coding, bioinformatics, and data mining, etc.

Thomas Bayes (1702–1761)

- Thomas Bayes was an English presbyterian minister whose mathematical writings earned him a place as a fellow of the Royal Society of London (1742)

- His friend Richard Price found a manuscript after his death and had it published.

- Bayes studied an "inverse probability" (i.e. inference) problem and laid the foundations of modern Bayesian statistics

- Probability : statements about observables given assumptions about unknown parameters

- Inverse probability : statements about unknown parameters given observed data values

The Fall and Rise of Bayesian Statistics

- Pierre Simon Laplace (1749-1827), French mathematician and astronomer who developed independently Bayesian ideas and applied them to mathematical astronomy and other fields. He refined inverse probability, acknowledging Bayes' work in a monograph in 1812.

- Although the Bayesian approach to inference was extensively developed (starting in the late eighteenth century) in 1920s - 1930s, in the twentieth century a completely different approach to inference, sometimes called Frequentist statistics, was developed and eventually came to dominate the field.

- Current huge popularity of Bayesian methods is due to fast computers and MCMC methods (since 1990s) : led to a dramatic rise even in analysis of complex models.

Review of Probability Basics (Aside)

- Probability is a mathematical representation for uncertainty.

- We assign probability to events :

An event A is a subset of the sample space Ω

- A probability measure is a function on events that satisfies:

- $P(A) \geq 0$ for all events A
- $P(\Omega) = 1$
- If $A_i \cap A_j = \emptyset$, then $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$

- From these properties we can derive others, e.g.

- If $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for events A and B

How to define a probability

1. Classical model-based approach:

- Probability $= \dfrac{\text{Number of outcomes}}{\text{Number of possible events}}$
- "All events are equally likely to happen."

2. Relative frequency (Frequentist) approach:

- Probability= long-run relative frequency = limit $\dfrac{\text{Number of events occurred}}{\text{Total number of trials}}$
- Based on repeated sampling/trials

3. Subjective probability (Bayesian) approach

- Assign some numerical value to some degrees of one's personal belief
- "All probabilities are conditional."

Examples

1. Model-based probability:

P("2" lands from a die) = ?

P("3" out of all cards) = ?

2. Coin tossing : H H T H T T T H H ⋯

P("Heads"): 1 2/2 2/3 ⋯

- unpredictable in short run, but stabilized in the long run

3. P(Rain today in Tampere) = $\begin{cases} ? & \text{for someone in U.S.} \\ ? & \text{for someone in Helsinki} \\ ? & \text{for someone in Tampere} \end{cases}$

P("H"| coin is fair) = ? P("H"| unlimited trials) = ?

Examples

1. Model-based probability:

   P("2" lands from a die) = 1/6

   P("3" out of all cards) = 4/52

2. Coin tossing :

2. Coin tossing :	H	H	T	H	T	T	T	H	H ⋯
P("Heads"):	1	2/2	2/3	3/4	3/5	3/6	3/7	4/8	5/9 ⋯

- unpredictable in short run, but stabilized in the long run

$$3.\ \ P(\text{Rain today in Tampere}) = \begin{cases} ? & \text{for someone in U.S.} \\ ? & \text{for someone in Helsinki} \\ ? & \text{for someone in Tampere} \end{cases}$$

P("H"| coin is fair) = 1/2   P("H"| unlimited trials) = 1/2

How Probability is Treated

- The frequentist approach treats probability as the long-run relative frequency of an event that can be repeated.

- Bayesian approach treats probability as the **relative plausibility** of an event.

- Consider the probability of getting a “head” when flipping a coin.

  - The frequentist would ask, “If we flipped the coin many times, what proportion of the flips would result in heads?”

  - The Bayesian would ask, “How likely is a “head” relative to the likelihood of a “tail”?

- The probabilities might be the same in each case, but the ways they are conceived of are different.

Subjective Probability

- In Bayesian statistics, probability represents degrees of belief (plausibility, confidence, credibility, certainty).

- All probabilities are conditional: the probability that you assign to something is always dependent on information that you have.

- The probability $P(E|H)$ is a number that measures your belief in the truth of event *E* given the knowledge (information) that *H* is true.

- It also obeys the probability (Kolmogorov) axioms:

1. $P(E|H) \geq 0$ 2. $P(H|H) = 1$

3. $P(E|H) + P(E^c|H) = 1$ 4. $P(E \cup F|H) = P(E|H) + P(F|H)$ when $E \cap F = \emptyset$

- Background information '*H*' is information that is assumed but not always stated. For simplicity, we often omit writing *H* explicitly; for $P(A|H)$ we usually write just P(*A*).

Conditional Probability and Total Probability

- The joint probability $P(\text{A}\cap\text{B})$ is the probability of observing both events A and B. (also denoted as $P(\text{AB})$ or $P(\text{A, B})$ for convenience)

- Conditional Probability: the conditional probability of B given A is

  $P(\text{B}|\text{A}) = P(\text{A}\cap\text{B})/ P(\text{A})$

  $\Longrightarrow P(\text{A}\cap\text{B}) = P(\text{B}|\text{A})\ P(\text{A})$: Multiplicative Law of Probability

- Events A and B are independent if any of the following holds:

  $P(\text{A}|\text{B}) = P(\text{A})$ or $P(\text{B}|\text{A}) = P(\text{B})$ or $P(\text{AB}) = P(\text{A})P(\text{B})$:

- The Law of Total Probability

$P(\text{A}) = P(\text{A, B}) + P(\text{A}, \sim\text{B}) = P(\text{A}\cap\text{B})+ P(\text{A}\cap\text{B}^c) = P(\text{A}|\text{B})P(\text{B})+ P(\text{A}|\text{B}^c)P(\text{B}^c)$

- Useful result: $P(\text{A, B}|\text{C}) = P(\text{A}|\text{B,C})\ P(\text{B}|\text{C})$

Bayes' Theorem

- Multiplicative law and the law of total probability lead to

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{B})} = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B},\,\mathrm{A}) + P(\mathrm{B},\,\mathrm{A}^c)}$$

$$= \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) + P(\mathrm{B}|\mathrm{A}^c)P(\mathrm{A}^c)}$$

- Theorem

Let A_1 , A_2,…, A_n be a set of mutually exclusive and exhaustive events. Then,

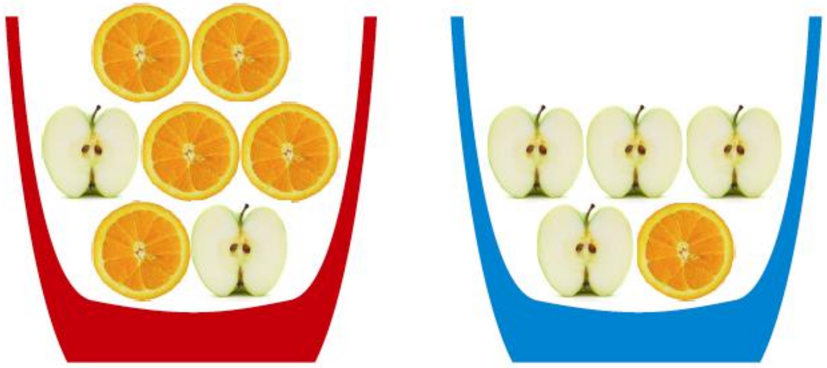
$$P(\mathrm{A}_i|\mathrm{B}) = \frac{P(\mathrm{A}_i)\,P(\mathrm{B}|\mathrm{A}_i)}{P(\mathrm{B})} = \frac{P(\mathrm{A}_i)\,P(\mathrm{B}|\mathrm{A}_i)}{\sum_{j=1}^{n} P(\mathrm{A}_j)\,P(\mathrm{B}|\mathrm{A}_j)}$$

Example : Discrete Probabilities

i) What is the probability of getting an orange if the bowl is red?

ii) What is the probability of the red bowl if the selected fruit is orange?

y \ x	🍏	🍊
🟥	2	5
🟦	4	1



- $p(x) = \sum_y p(x, y)$
- $p(x, y) = p(x|y)\, p(y)$

i) What is the probability of getting an orange if the bowl is red?

ii) What is the probability of the red bowl if the selected fruit is orange?

x / y			
	2	5	7
	4	1	5
	6	6	12

- Let $P(\text{O}|\text{R}) \equiv P(x\text{=orange}\,|y\text{= red})$

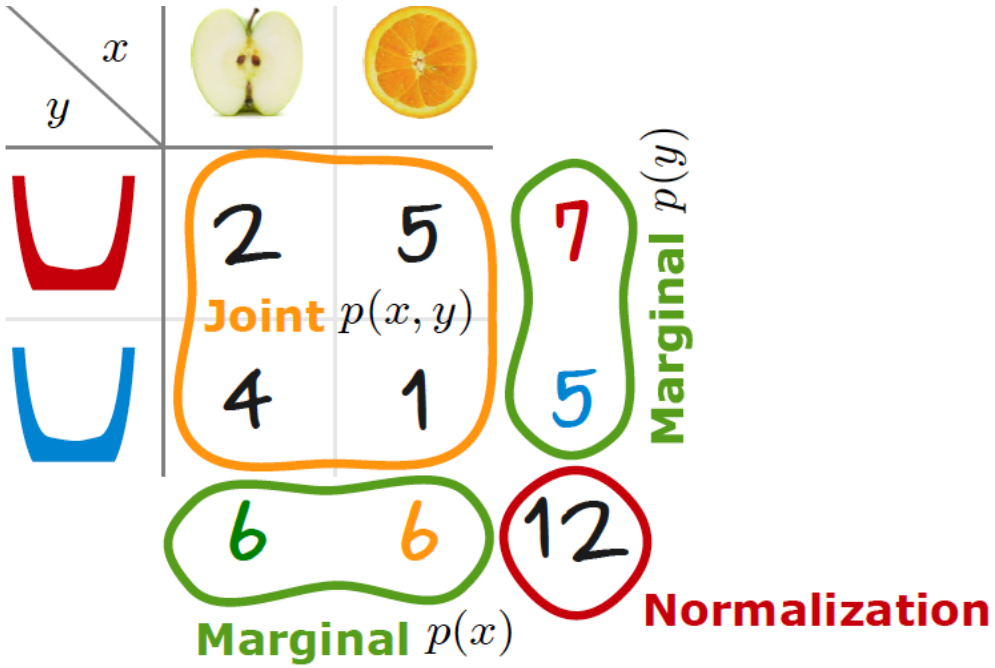
$$P(\text{O}|\text{R}) = \frac{P(\text{R, O})}{P(\text{R})}$$

$$P(\text{R}|\text{O}) = \frac{P(\text{R, O})}{P(\text{O})}$$

- $P(\text{R}|\text{O}) = \dfrac{P(\text{R, O})}{P(\text{O})} = \dfrac{P(\text{O}\,|\,\text{R})P(\text{R})}{P(\text{O})}$

i) What is the probability of getting an orange if the bowl is red?

ii) What is the probability of the red bowl if the selected fruit is orange?



- Let $P(\text{O}|\text{R}) \equiv P(x\text{=orange}\,|y\text{= red})$

$$P(\text{O}|\text{R}) = \frac{P(\text{R, O})}{P(\text{R})} = \frac{5/12}{7/12} = \frac{5}{7}$$

$$P(\text{R}|\text{O}) = \frac{P(\text{R, O})}{P(\text{O})} = \frac{5/12}{6/12} = \frac{5}{6}$$

- $P(\text{R}|\text{O}) = \dfrac{P(\text{R, O})}{P(\text{O})} = \dfrac{P(\text{O}|\text{R})P(\text{R})}{P(\text{O})} = \dfrac{5/7 \text{ x } 7/12}{6/12} = \dfrac{5}{6}$

Medical Tests and Bayes’ Theorem

- Example: Suppose that you are worried that you might have a rare disease.

Assume that the probability of having such disease in general is about 0.0001. You decide to get tested, and suppose that the testing methods for this disease are correct 90 percent of the time. However, it also gives false positive results with 0.01 probability.

- What is the probability that you really have the disease when the test gives positive result?

$P[\text{Disease}] = 0.0001$ $P[+|D\ ]= 0.9$ $P[+|\text{No D}]= 0.01$

Note: $P[\ +\ ] = P[+, D] + P[+, \text{No D}]$

Medical Tests and Bayes' Theorem

- Example: Suppose that you are worried that you might have a rare disease.

Assume that the probability of having such disease in general is about 0.0001. You decide to get tested, and suppose that the testing methods for this disease are correct 90 percent of the time. However, it also give false positive results with 0.01 probability.

- What is the probability that you really have the disease when the test gives positive result?

P[Disease] = 0.0001 P[+|D]= 0.9 P[+|No D]= 0.01

$$P[\ D\ |\ +\ ] = \frac{P[D\ ,+\ ]}{P[\ +\ ]} = \frac{P[+|D\ ]P[D\ ]}{P[+|D]P[D]+P[+|\sim D]P[\sim D]}$$

$$= \frac{0.9 \text{ x } 0.0001}{0.9\text{x } 0.0001+0.01\text{x } 0.9999} \approx 0.08$$

Example: Icy Roads

- Inspector Smith is waiting for Holmes and Watson who are both late for an appointment. Smith is worried that if the roads are icy, one or both of them may have crashed his car.

$P(\text{Icy}) = 0.7$ $P(\text{Not icy}) = 0.3$

$P(\text{ One crashes}|\text{icy}) = 0.8$ $P(\text{ One crashes}|\text{ not icy}) = 0.1$

i) What is the probability that Holmes crashes his car?

ii) Suddenly Smith learns that Watson has crashed his car.

What is the probability that Holmes crashes his car given the information that Watson has crashed?

i) P(Holmes crashes)= P(Holmes crashes, icy)+ P(Holmes crashes, not icy)

= P(Holmes crashes|icy) P(Icy) + P(Holmes crashes|not icy)P(Not icy)

= 0.8x0.7 + 0.1x0.3 = 0.59

ii) P(Holmes crashes | Watson crashed)

= P(Holmes crashes, Icy | Watson crashed) + P(Holmes crashes, Not icy | Watson crashed)

= P(Holmes crashes | Icy, Watson crashed) P(Icy | Watson crashed)

+ P(Holmes crashes| Not icy, Watson crashed) P(Not icy|Watson crashed)

= P(Holmes crashes | Icy) P(Icy | Watson crashed)

+ P(Holmes crashes | Not icy) P(Not icy |Watson crashed)

: P(A, B|C) = P(A|B,C) P(B|C) and Holmes' crash is conditionally independent of Watson's crash given the icy conditions.

- P(Holmes crashes | Watson crashed)

= P(Holmes crashes | Icy) P(Icy | Watson crashed)

+ P(Holmes crashes | Not icy) P(Not icy |Watson crashed)

= 0.8 x 0.95 + 0.1 x 0.05 = 0.765

- P(Icy|Watson crashed) = $\dfrac{\text{P(Watson crashed| Icy)xP(Icy)}}{\text{P(W crashed)}} = \dfrac{0.8\text{x}0.7}{0.59} = 0.95$

1. P(Holmes crashes) = 0.5 without info. on road conditions
2. P(Holmes crashes) = 0.59 when the road is icy.
3. P(Holmes crashes | Watson crashed) = 0.765

Example: multiple medical tests

- A patient goes to see a doctor due to some discomfort.

Doctor believes that he may have a disease A.

$$\text{Let } \begin{cases} \theta = 1 & \text{if having disease A} \\ \theta = 0 & \text{if not having disease A} \end{cases}$$

- Doctor’s past experience: P[$\theta = 1$] = 0.7

- The patient undertakes an examination Y :

P[Y= 1 | $\theta = 0$] = 0.40 : positive result without disease

P[Y= 1 | $\theta = 1$] = 0.95 : positive result with disease

- Data: $Y = 1$ (positive result)

Example (cont'd)

i) What is the probability that he has the disease A given '+' test?

$$\text{P}[\,\theta = 1 \mid Y = 1] = \frac{\text{P}[\theta=1,\, Y=1]}{\text{P}[Y=1]} = \frac{\text{P}[Y = 1|\theta=1]\text{P}[\theta=1]}{\text{P}[Y=1]} = \frac{0.95 \text{ x } 0.7}{0.95\text{x } 0.7+0.40\text{x } 0.3}$$

$$= \frac{0.665}{0.785} \approx 0.847$$

- $\text{P}[Y = 1] = \text{P}[Y = 1, \theta = 1] + \text{P}[Y = 1, \theta = 0]$ (law of total prob.)

  $= \text{P}[Y = 1|\theta = 1]\ \text{P}[\theta = 1] + \text{P}[Y = 1|\theta = 0]\ \text{P}[\theta = 0]$

  $= 0.95 \text{ x } 0.7 + 0.4 \text{ x } 0.3$

  $= 0.665 + 0.120 = 0.785$

- Suppose that there is a more reliable test 'W' : $\begin{cases} \text{P}[W = 1 \mid \theta = 0] = 0.04 \\ \text{P}[W = 1 \mid \theta = 1] = 0.99 \end{cases}$

- 2nd data: $W = 0$

ii) What is the probability of the patient being ill given two information from 2 (independent) tests?

$$\text{P}[\theta = 1 \mid Y = 1, W = 0] = \frac{\text{P}[\theta = 1, W = 0 \mid Y = 1]}{\text{P}[W = 0 \mid Y = 1]}$$

$$\begin{aligned} \text{P}[W = 0, \theta = 1 \mid Y = 1] &= \text{P}[W = 0 \mid \theta = 1, Y = 1]\ \text{P}[\theta = 1 \mid Y = 1] \\ &= \text{P}[W = 0 \mid \theta = 1]\ \text{P}[\theta = 1 \mid Y = 1] \\ &= 0.01 \text{ x } 0.847 \approx 0.008 \end{aligned}$$

Sequential Update of Information

- P[W = $0|Y = 1$]= P[W = 0, $\theta = 1|Y = 1$] + P[W = 0, $\theta = 0|Y = 1$]

  = 0.008 + P[W = $0|\theta = 0$] P[$\theta = 0|Y = 1$]

  = 0.008 + 0.96(1 − 0.847) = 0.008+0.147 = 0.155

P[$\theta = 1$ | Y= 1, W = 0] = 0.008/0.155= 0.052

- Doctor’s findings:

  - P[$\theta = 1$] = 0.7 before tests Y & W
  - P[$\theta = 1$| Y= 1] = 0.847 after test Y
  - P[$\theta = 1$| Y= 1, $W = 0$] = 0.052 after test W

Recap : Bayes' theorem

- Bayes' theorem is a formula for learning.

- Suppose we have an initial or prior belief about the truth of A. We observe some data D. Then, we calculate our revised or posterior belief about the truth of A, in the light of the new data D, using Bayes' theorem.

$$P(\mathrm{A}|\mathrm{D}) = \frac{P(\mathrm{A}\cap\mathrm{D})}{P(\mathrm{D})} = \frac{P(\mathrm{D}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{D})}$$

- P(A) encodes our prior distribution.
- P(D|A) forms the model for data. P(D) is the marginal probability of data.
- P(A|D) presents our posterior belief in A after having observed D.

- Bayesian analyses balance prior information and information from data.

Statistics Using Bayes' Rule

- We now consider inference about parameters, based on data.
- Generically denote an unknown parameter of interest as θ and data as D.
- Our probability model for the data, given a value of θ is denoted $P(\mathrm{D}|\theta)$.
- Our model for our prior knowledge about θ is denoted $P(\theta)$.
- We seek to make formal probability statements about θ given some observed data : $P(\theta|\mathrm{D})$

$$P(\theta|\mathrm{D}) = \frac{P(\mathrm{D}|\theta)P(\theta)}{P(\mathrm{D})}$$

Bayesian Method for Inference

1. Prior

Specify the prior distribution: $[\theta]$, $f(\theta)$

which expresses our knowledge about θ prior to observing the data.

2. Likelihood

Model a set of observations with a probability distribution (expressed in the form of the likelihood function) with unknown parameter(s): $[x|\theta]$, $f(x|\theta)$

3. Posterior

Apply Bayes' theorem to derive posterior distribution : $[\theta|x]$, $f(\theta|x)$

which expresses all that is known about θ after observing the data.

4. Inference

Derive appropriate inference statements from the posterior distribution: e.g. point / interval estimates, probabilities of specified hypotheses.

Two main approaches to statistical inference

i) Frequentist/conventional/classical approach

- Parameters are fixed but unknown quantities
- Data are drawn from a distribution of known form but with an unknown parameter. Often this distribution arises from explicit randomization.
- Inferences regard the data as random and repeated sampling is assumed.

ii) Bayesian approach

- Parameters (unknown quantities) are random variables
- Probability distributions are assumed for the unknown parameters and for the observations (i.e. both parameters and observations are random quantities).
- Inferences are based on the prior distribution and the observed data.

Why do people use classical methods?

- If there is no prior information available about the parameter(s).

- If they prefer “cookbook”-type formulas with little input from the scientists /researchers.

- Bayesian methods require a bit more mathematical formalism.

- Historically (**but not now**) realistic Bayesian analyses had been infeasible

due to a lack of computing power.

- Many methods were developed in the context of controlled experiments. Then, the parameters of interest can be regarded as truly fixed quantities.

Why use Bayesian methods?

- We can specifically incorporate previous knowledge (and expert judgement) we have about a parameter of interest.
- To logically update our knowledge about the parameter after observing data.
- Offers flexibility in statistical modelling: e.g. Highly nonlinear models with many parameters can be analyzed.
- Can handle "nuisance" parameters that pose problems for frequentist inference.
- Does not rely on large sample asymptotics, but gives valid inference also for small sample sizes.