MA930 Data Analysis & Machine Learning

Lecture 4:Bayesian Statistics

Haoran Ni

Recap from last class

- Focus on type-I error rate: i.e. $\alpha = \boldsymbol{P}(\text{reject null}|\text{ null true})$, compare with p value
- Involves reject/do not reject the null hypothesis given available evidence from an experiment
- Many different types of hypothesis test, applicable in different scenarios and with different datasets
- Widely applied, particularly in medical/biological literature

Recap from last class

Frequentist	Bayesian
Parameter values treated as *fixed* quantities (i.e. true value exists but is unknown, and we wish to estimate it)	Parameters are treated as random variables (i.e. they follow a probability distribution)
The frequentist interpretation of probability corresponds to the uncertainty in an outcome of an experiment each time, if the experiment was repeated an infinite number of times	The Bayesian interpretation of probability corresponds to how a degree of belief in a proposition changes due to available evidence
Confidence intervals	**Credible** intervals

Outline

- Bayesian basics
- Likelihood functions
- Conjugate priors
- Credible intervals
- Epidemiological example

Bayesian basics

- Bayes' theorem/rule: $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$
- Apply this to data **D** and a prior/hypothesis **H**
- This allows us to update our beliefs using information from an experiment

- $P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$, or in other words: $P(H|D) \propto P(D|H)\,P(H)$

- $P(H)$ is the prior belief about the hypothesis
- $P(D|H)$ is called the "likelihood" = probability of data given the hypothesis
- $P(D)$ is the probability of the data – a normalisation constant
- $P(H|D)$ is called the "posterior" – this is what you want to know

Likelihood functions

- For independent data points, $D = \{D_k\}$, and the likelihood function takes the form of a product
- $P(H|D) \propto P(D|H)\, P(H) = \prod_k (P(D_k|H))\, P(H)$

- If there are two hypotheses, H_1 and H_2, then we can normalise this as follows:

$$P(\mathrm{H}_1|D) = \frac{1}{M}\prod\nolimits_k (P(D_k|\mathrm{H}_1))\, P(\mathrm{H}_1)$$

$$P(\mathrm{H}_2|D) = \frac{1}{M}\prod\nolimits_k (P(D_k|\mathrm{H}_2))\, P(\mathrm{H}_2)$$

Where $M = \prod_k (P(D_k|\mathrm{H}_1))\, P(\mathrm{H}_1) + \prod_k (P(D_k|\mathrm{H}_2))\, P(\mathrm{H}_2)$

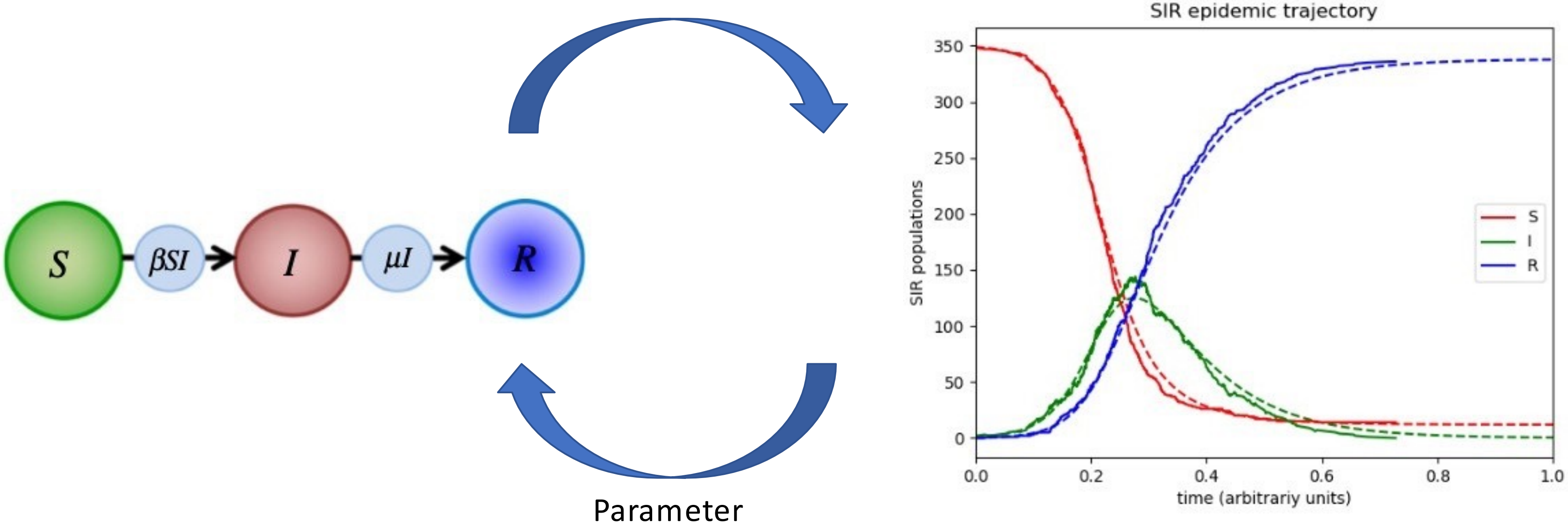
Likelihood functions

Example 4.1. Consider a game in which one of two coins is used. One is fair, and the other one is biased for heads (with probability 0.7 of getting a head per throw). From experience, we know that the biased coin is used in the game 20% of the time. A coin is chosen, and the following sequence is obtained: HTHHHHTH. We want to know which coin was used.

i) What is the prior distribution?
ii) What are the likelihoods following the first flip?
iii) What is the posterior distribution following the first flip?
iv) What is the posterior distribution at the end of the sequence?
v) What would the posterior distribution be at the end if there was no prior knowledge?

Likelihood functions

- Likelihood functions are often used to estimate parameters of models
- Parameter inference is essentially the inverse of model simulation



Simulation

Parameter inference

Likelihood functions

- One approach is maximum likelihood estimation, which involves finding the parameter values that maximise the likelihood function (or, equivalently, the log-likelihood function; often better computationally when likelihoods are small).
- The full Bayesian approach is to construct the posterior, as described previously.
- $P(\theta|D) = \frac{P(D|\theta)\, P(\theta)}{P(D)}$, or in other words:$P(\theta|D) \propto P(D|\theta)\, P(\theta)$ [posterior proportional to likelihood x prior]
- When estimating k multiple parameters, obtain a likelihood surface (in k dimensions) and the posterior is a “joint posterior”

Likelihood functions

Exercise 4.1. Maximum likelihood estimation

A lightbulb manufacturer wishes to produce an estimate for the mean lifetime of their bulbs, based on observations $X_1, X_2, \ldots, X_n$.The lifetime of a bulb is drawn from an exponential distribution with rate parameter .

i) Show that the maximum likelihood estimate (MLE) corresponds to a lifetime of $1/\bar{X}$. (i.e. write down the likelihood, and then – by differentiating and setting equal to zero – find the value of λ that maximises it; how do you know it's a maximum?)

ii) In a specific case, $n = 1000$, $\bar{X} = 368.15$ days. Plot the log-likelihood function and find the MLE.

Likelihood functions

Exercise 4.2. Bayesian posterior distribution

In the lightbulb example from the previous slide, find an analytic expression for the posterior for λ, assuming a Gamma distributed prior with shape parameter k and scale parameter θ.

Verify that the posterior is also a gamma distribution.

Reminder: The PDF of a gamma distribution is given by: $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp(-\frac{x}{\theta})$

Conjugate priors

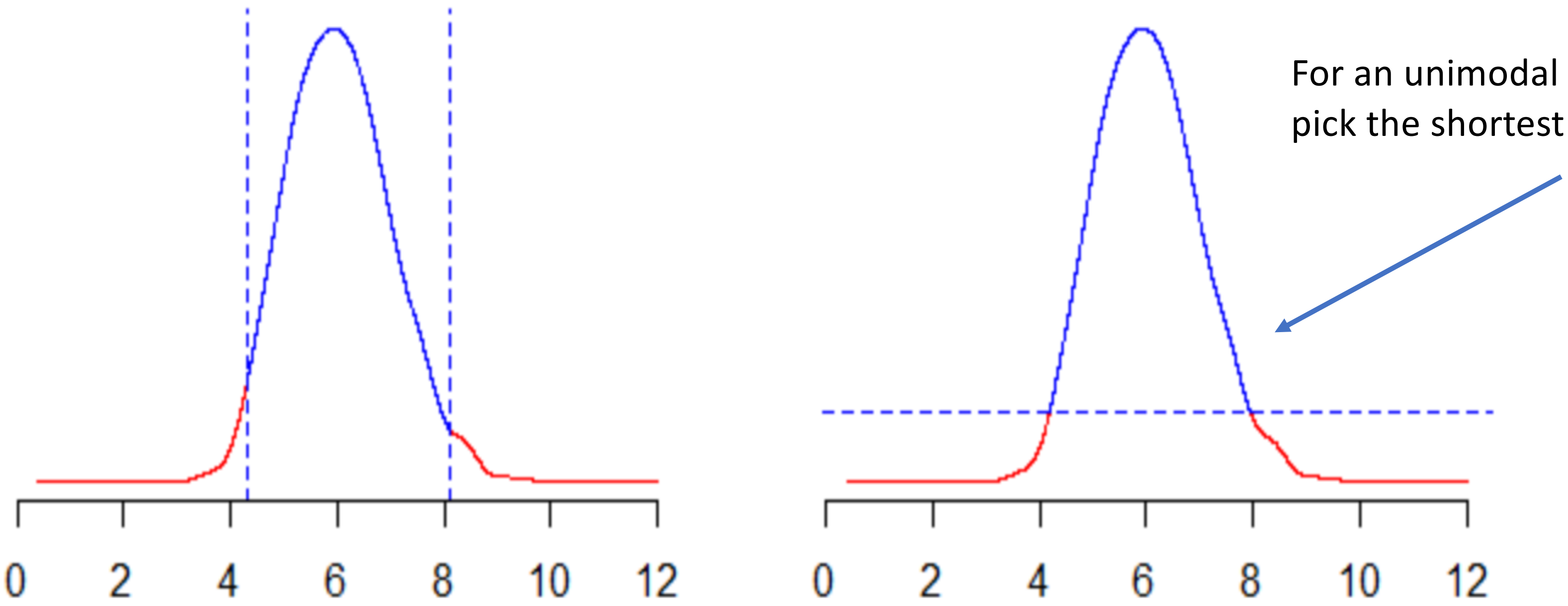
- In the previous example, a specific choice of prior distribution allowed us to evaluate the posterior analytically
- This choice of prior distribution is called a *conjugate prior.*

- Note that: When estimating the parameter of an exponential distribution using a Gamma distributed prior (shape, scale), then the posterior is also Gamma distributed with shape and scale parameters that can be calculated analytically.

Credible intervals

- After calculating the posterior, the ***central* 95% credible interval** (or 95% quantile interval/95% equal- tailed interval) is the central interval in which 95% of the posterior lies (i.e. 2.5% of the posterior lies below it, and 2.5% lies above it). The parameter has probability 0.95 of lying within it
- The ***highest density* 95% credible interval** is the smallest possible interval containing 95% of the posterior

Credible intervals


Central 95% CrI = [4.2, 8.1]

Highest density 95% CrI = [4.1, 8.0]

For an unimodal distribution, pick the shortest possible 95% CrI

Credible intervals

Exercise 4.3. Bayesian posterior distribution

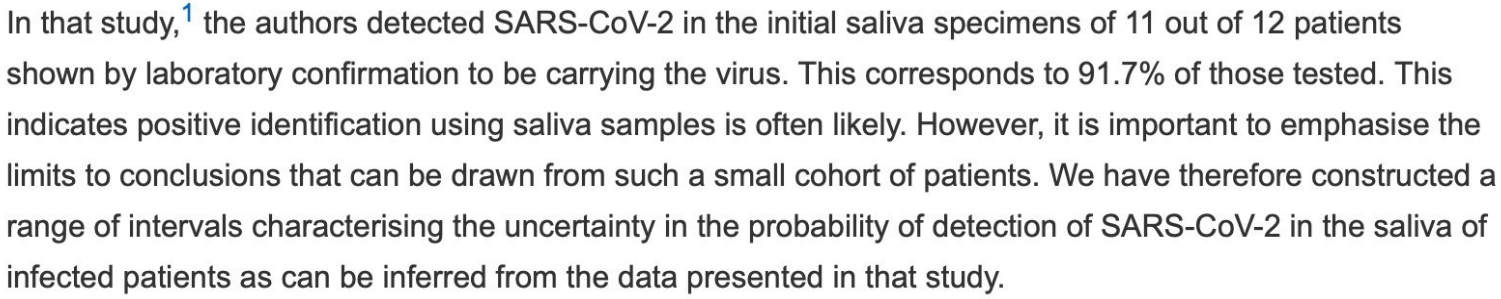
Plot the posterior distributions corresponding to the lightbulb example with $n = 1000$, $\bar{X} = 368.15$ days, with a Gamma prior with $k = 0.0001$ and $\theta = 1000$. Also plot the prior distribution and calculate a central 95% credible interval for λ.

Now repeat the above, but without using your knowledge about conjugate priors.

Hint: to avoid numerical errors, it may be necessary to work with log's when performing the calculations. Note that you can add or subtract any constant value from the log posterior to give the same result when the (normalised) posterior is calculated

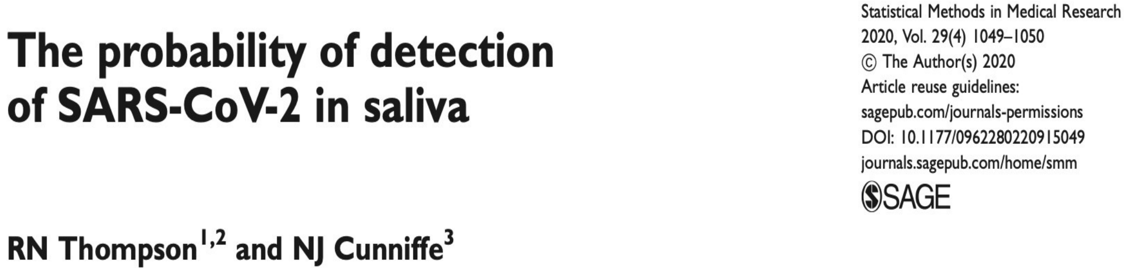
Credible intervals

Exercise 4.4. Credible intervals: Saliva tests for COVID-19



In that study,[1] the authors detected SARS-CoV-2 in the initial saliva specimens of 11 out of 12 patients shown by laboratory confirmation to be carrying the virus. This corresponds to 91.7% of those tested. This indicates positive identification using saliva samples is often likely. However, it is important to emphasise the limits to conclusions that can be drawn from such a small cohort of patients. We have therefore constructed a range of intervals characterising the uncertainty in the probability of detection of SARS-CoV-2 in the saliva of infected patients as can be inferred from the data presented in that study.

Assuming a binomial distribution for the likelihood, and a uniform prior on [0,1] for detection probability p, construct both central and highest density 95% credible intervals for p. Check your answer in the publication screenshot below.



The probability of detection of SARS-CoV-2 in saliva

RN Thompson[1,2] and NJ Cunniffe[3]

Statistical Methods in Medical Research
2020, Vol. 29(4) 1049–1050
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220915049
journals.sagepub.com/home/smm
SAGE

Additional questions

Exercise 4.5. Coin tossing – inferring bias

A coin is flipped N times with n heads. We want to infer the probability that a coin flip gives a head (p).

i) What is the likelihood function?

ii) For a coin with a bias of 0.7 towards heads, write code to generate 20 coin flips. Calculate posteriors corresponding to two priors: one uniform and the other proportional to $p(1-p)$.

Exercise 4.6. Conjugate priors: inferring bias in coin flips

The conjugate prior to the binomial distribution (likelihood) is the beta distribution.

i) Suppose we start with a beta-distributed prior with parameters α and β. Show that if a coin is flipped $h + t$ times with h heads and t tails, the posterior for p (the probability that a coin toss gives a head) is a beta distribution with $\alpha' = \alpha + h$ times and $\beta' = \beta + t$

i) For a coin with a 0.7 heads bias, generate 50 coin tosses. Choose a beta prior with $\alpha = 2$ and $\beta = 2$. Plot the prior and posterior on the same axes, and then calculate central and highest-density 95% credible intervals for p. Indicate the credible intervals on the same axes too.

Additional questions

Exercise 4.7. Posterior for the normal distribution

For some choice of parameters, generate $n = 5$ normally distributed random numbers.

i) Using a flat prior, calculate and plot the posterior density. [n.b. this is a twodimensional density, so you will need a 2d grid of data points]

ii) Plot the marginal posteriors