Quick introduction to simulation and sampling

Azim Ansari
ansari@well.ox.ac.uk
University of Oxford

15/11/2019

We have some data and would like to model it:

- ▶ We assume a model $p(data|\theta)$ for how the data was generated,

We have some data and would like to model it:

- ▶ We assume a model $p(data|\theta)$ for how the data was generated,
- ▶ We set a prior $p(\theta)$ on the parameters,

We have some data and would like to model it:

► We assume a model $p(data|\theta)$ for how the data was generated,

► We set a prior $p(\theta)$ on the parameters,

Then by Bayes' Theorem:

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{\int_\theta p(data|\theta)p(\theta)d\theta}$$

UNIVERSITY OF
OXFORD

We have some data and would like to model it:

- ▶ We assume a model $p(data|\theta)$ for how the data was generated,
- ▶ We set a prior $p(\theta)$ on the parameters,

Then by Bayes' Theorem:

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{\int_\theta p(data|\theta)p(\theta)d\theta}$$

$$= \frac{p(data|\theta)p(\theta)}{P(data)}$$

We have measured the number of de novo mutations in the genome of 10 children (by comparing their genome to their parents genome).

Data:
42, 49, 40, 36, 50, 37, 39, 43, 54, 55

Model the data and answer this questions:

► What is the probability that mutation rate > 45 mutations per genome?

Have a go and see if you can do it on a piece of paper. Just follow the steps mentioned previously.

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{likelihood:} \quad j|\lambda \sim \text{Poisson}(\lambda)$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{likelihood:} \quad j|\lambda \sim \text{Poisson}(\lambda)$$

$$\text{likelihood:} \quad P(j|\lambda) = \frac{e^{-\lambda}\lambda^j}{j!}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{likelihood:} \quad j|\lambda \sim \text{Poisson}(\lambda)$$

$$\text{likelihood:} \quad P(j|\lambda) = \frac{e^{-\lambda}\lambda^j}{j!}$$

$$\text{prior:} \quad \lambda \sim \text{Gamma}(a, b)$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{likelihood: } j|\lambda \sim \text{Poisson}(\lambda)$$

$$\text{likelihood: } P(j|\lambda) = \frac{e^{-\lambda}\lambda^j}{j!}$$

$$\text{prior: } \lambda \sim \text{Gamma}(a, b)$$

$$\text{prior: } p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{likelihood: } j|\lambda \sim \text{Poisson}(\lambda)$$

$$\text{likelihood: } P(j|\lambda) = \frac{e^{-\lambda}\lambda^j}{j!}$$

$$\text{prior: } \lambda \sim \text{Gamma}(a, b)$$

$$\text{prior: } p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

$$\text{posterior: } p(\lambda|j) = \frac{P(j|\lambda)p(\lambda)}{P(j)}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{posterior:} \quad p(\lambda|j) = \frac{P(j|\lambda)p(\lambda)}{P(j)}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{posterior:} \quad p(\lambda|j) = \frac{P(j|\lambda)p(\lambda)}{P(j)}$$
$$= \frac{\frac{e^{-\lambda}\lambda^j}{j!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j)}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\text{posterior: } p(\lambda|j) = \frac{P(j|\lambda)p(\lambda)}{P(j)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^j}{j!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j)}$$

$$= \left[\frac{b^a}{j!\Gamma(a)P(j)}\right]\lambda^{(a+j)-1}e^{-(b+1)\lambda}$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$
\begin{aligned}
\text{posterior:} \quad p(\lambda|j) &= \frac{P(j|\lambda)p(\lambda)}{P(j)} \\
&= \frac{\frac{e^{-\lambda}\lambda^j}{j!} \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j)} \\
&= \left[\frac{b^a}{j!\Gamma(a)P(j)}\right] \lambda^{(a+j)-1}e^{-(b+1)\lambda} \\
&= C \, \lambda^{(a+j)-1}e^{-(b+1)\lambda}
\end{aligned}
$$

We assume that the data is generated according to a Poisson distribution. First look at a single observation:

$$\begin{aligned} \text{posterior:} \quad p(\lambda|j) &= \frac{P(j|\lambda)p(\lambda)}{P(j)} \\ &= \frac{\frac{e^{-\lambda}\lambda^j}{j!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j)} \\ &= \left[\frac{b^a}{j!\Gamma(a)P(j)}\right]\lambda^{(a+j)-1}e^{-(b+1)\lambda} \\ &= C\,\lambda^{(a+j)-1}e^{-(b+1)\lambda} \end{aligned}$$

This is a $\text{Gamma}(a + j, b + 1)$ PDF. So we have the posterior distribution of the model parameter.

OXFORD

Now consider m i.i.d. observation:

$$j_1, \ldots, j_m \overset{iid}{\sim} \text{Poisson}(\lambda)$$

Now consider m i.i.d. observation:

$$j_1, \ldots, j_m \overset{iid}{\sim} \text{Poisson}(\lambda)$$
$$P(j_1, ..., j_m | \lambda) = P(j_1 | \lambda) \ldots P(j_m | \lambda)$$

Now consider m i.i.d. observation:

$$j_1, \ldots, j_m \overset{iid}{\sim} \text{Poisson}(\lambda)$$
$$P(j_1, ..., j_m|\lambda) = P(j_1|\lambda) \ldots P(j_m|\lambda)$$
$$\text{likelihood:} \quad P(j_1, \ldots, j_m|\lambda) = \frac{e^{-\lambda}\lambda_1^j}{j_1!} \ldots \frac{e^{-\lambda}\lambda_m^j}{j_m!}$$

Now consider m i.i.d. observation:

$$j_1, \ldots, j_m \overset{iid}{\sim} \text{Poisson}(\lambda)$$
$$P(j_1, ..., j_m | \lambda) = P(j_1 | \lambda) \ldots P(j_m | \lambda)$$
$$\text{likelihood:} \quad P(j_1, \ldots, j_m | \lambda) = \frac{e^{-\lambda} \lambda_1^j}{j_1!} \ldots \frac{e^{-\lambda} \lambda_m^j}{j_m!}$$
$$\text{prior:} \quad \lambda \sim \text{Gamma}(a, b)$$
$$\text{prior:} \quad p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

Now consider m i.i.d. observation:

$$j_1, \ldots, j_m \overset{iid}{\sim} \text{Poisson}(\lambda)$$

$$P(j_1, \ldots, j_m | \lambda) = P(j_1 | \lambda) \ldots P(j_m | \lambda)$$

$$\text{likelihood:} \quad P(j_1, \ldots, j_m | \lambda) = \frac{e^{-\lambda} \lambda_1^j}{j_1!} \ldots \frac{e^{-\lambda} \lambda_m^j}{j_m!}$$

$$\text{prior:} \quad \lambda \sim \text{Gamma}(a, b)$$

$$\text{prior:} \quad p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

$$\text{posterior:} \quad p(\lambda | j_1, \ldots, j_m) = \frac{P(j_1, \ldots, j_m | \lambda) p(\lambda)}{P(j)}$$

$$p(\lambda|j_1, \ldots, j_m) = \frac{P(j_1, \ldots, j_m|\lambda)p(\lambda)}{P(j_1, \ldots, j_m)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{j_1}}{j_1!} \cdots \frac{e^{-\lambda}\lambda^{j_m}}{j_m!} \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j_1, \ldots, j_m)}$$

$$p(\lambda|j_1,\ldots,j_m) = \frac{P(j_1,\ldots,j_m|\lambda)p(\lambda)}{P(j_1,\ldots,j_m)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{j_1}}{j_1!}\cdots\frac{e^{-\lambda}\lambda^{j_m}}{j_m!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j_1,\ldots,j_m)}$$

$$= \left[\frac{b^a}{j_1!\ldots j_m!\Gamma(a)P(j_1,\ldots,j_m)}\right]\times$$

$$\lambda^{(a+j_1+\cdots+j_m)-1}e^{-(b+m)\lambda}$$

$$p(\lambda|j_1,\ldots,j_m) = \frac{P(j_1,\ldots,j_m|\lambda)p(\lambda)}{P(j_1,\ldots,j_m)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{j_1}}{j_1!}\cdots\frac{e^{-\lambda}\lambda^{j_m}}{j_m!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j_1,\ldots,j_m)}$$

$$= \left[\frac{b^a}{j_1!\ldots j_m!\Gamma(a)P(j_1,\ldots,j_m)}\right]\times$$
$$\lambda^{(a+j_1+\cdots+j_m)-1}e^{-(b+m)\lambda}$$

$$= C\,\lambda^{(a+j_1+\cdots+j_m)-1}e^{-(b+m)\lambda}$$

$$p(\lambda|j_1, \ldots, j_m) = \frac{P(j_1, \ldots, j_m|\lambda)p(\lambda)}{P(j_1, \ldots, j_m)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{j_1}}{j_1!} \cdots \frac{e^{-\lambda}\lambda^{j_m}}{j_m!} \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{P(j_1, \ldots, j_m)}$$

$$= \left[\frac{b^a}{j_1! \ldots j_m!\Gamma(a)P(j_1, \ldots, j_m)} \right] \times$$
$$\lambda^{(a+j_1+\cdots+j_m)-1}e^{-(b+m)\lambda}$$

$$= C \, \lambda^{(a+j_1+\cdots+j_m)-1}e^{-(b+m)\lambda}$$

This is a Gamma$(a + j_1 + \cdots + j_m, b + m)$ PDF. So we have the posterior distribution of the model parameter.

UNIVERSITY OF
OXFORD

Now, let's calculate the probability that the de novo mutation rate > 45 mutations per genome:

$$P(\lambda > 45|j_1, \ldots, j_m) = \int_{45}^{\infty} C \; \lambda^{(a+\sum_1^m j_i)-1} e^{-(b+m)\lambda} d\lambda$$

$$\text{Where } C = \left[\frac{b^a}{j_1! \ldots j_m! \Gamma(a) P(j_1, \ldots, j_m)} \right]$$

In this instance we do not need to do integration as there is an R function that gives us the Gamma CDF. You just need to use 1-pgamma(45,shape=$a + j_1 + \cdots + j_m$, rate=$b + m$).

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.
2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generted the data.

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.

2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generted the data.

3. We Decide on the prior distribution of the model parameters.

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.
2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generted the data.
3. We Decide on the prior distribution of the model parameters.
4. Using Bayes' Theorem we have the posterior distribution for the parameters of the model.

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.

2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generated the data.

3. We Decide on the prior distribution of the model parameters.

4. Using Bayes' Theorem we have the posterior distribution for the parameters of the model.

5. Many realistic models are too complex for analytical solutions.

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.

2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generted the data.

3. We Decide on the prior distribution of the model parameters.

4. Using Bayes' Theorem we have the posterior distribution for the parameters of the model.

5. Many realistic models are too complex for analytical solutions.

6. Describe the model (distribution) using a large number of samples generated from it.

UNIVERSITY OF
OXFORD

1. We Make an assumption on how the data was generated, i.e. what probability model is likely to fit the data best.

2. I.e. we assume we know how the data was generated and would like say something about the parameters of the model that generted the data.

3. We Decide on the prior distribution of the model parameters.

4. Using Bayes' Theorem we have the posterior distribution for the parameters of the model.

5. Many realistic models are too complex for analytical solutions.

6. Describe the model (distribution) using a large number of samples generated from it.

7. Monte Carlo simply refers to simulating from a distribution and using the samples to estimate whatever we

Download the height data and read into R. Height data ignoring gender.

Your task:

1. What is the mean height?

2. What is the variance of the height?

3. What is the probability that a randomly chosen individual has a heigth of < 165 cm?

4. What is the probability that a randomly chosen individual has a height in the range $[180, 190]$ cm?

Go to R.

Simulation is a powerful technique even for simple problems.
The Monty Hall problem

1. There is a game show called Let's make a deal, hosted by Monty Hall.

2. Contestant chooses one of three closed doors, two of which have a goat behind them and one of which has a car.

3. Monty, who knows where the car is, then opens one of the two remaining doors.

4. The door he opens always has a goat behind it (he never reveals the car!).

5. Monty then offers the contestant the option of switching to the other unopened door.

6. If the contestant's goal is to get the car, should she switch doors?

▶ What do you think? Does it make a difference if the player swaps or not?

▶ We could solve this problem by using probabilistic arguments. Or we could simulate and see what the answer is.
  ▶ Simulate the two strategies separately.
  ▶ First strategy: do not swap
  ▶ Second strategy: always swap

Simulation procedure for the switching strategy

```
1 Repeat N times
2       randomly pick where the car is.
3       receive the players choice of door.
4       pick Montys door (cant be the players
            door or the car door).
5       player switches door.
6       determine the result of the game
7 Estimate the probability of winning using the N
      simulations.
```

Go to R.

The law of large numbers describes the behavior of the sample mean of i.i.d. r.v.s as the sample size grows. Assume we have i.i.d. $X_1, X_2, X_3, ...$ with finite mean μ and finite variance σ^2. For all positive integers n, let

$$\bar{X}_n = \frac{X_1 + ... + X_n}{n} \tag{1}$$

be the sample mean of X_1 through X_n.

Strong law of large numbers:

The sample mean $\bar{X}_n$ converges to the true mean μ as $n \to \infty$, with probability 1. In other words, the event $\bar{X}_n \to \mu$ has probability 1.

The law of large numbers is essential for simulations, statistics, and science. Consider generating "data" from a large number of independent replications of an experiment, performed either by computer simulation or in the real world. Every time we use the proportion of times that something happened as an approximation to its probability, we are implicitly appealing to LLN. Every time we use the average value in the replications of some quantity to approximate its theoretical average, we are implicitly appealing to LLN.

UNIVERSITY OF
OXFORD

- ▶ Samples from a distribution can be used to describe the distribution.
  - ▶ Estimate the mean or variance of the r.v.
  - ▶ What is the probability that the r.v. takes a value within a specific range.
- ▶ The main question is how do we generate $X_1, X_2, ..., X_n$ from a distribution.
- ▶ This is a huge field and I am only going to talk about Markov Chain Monte Carlo algorithm for sampling from a distribution.

UNIVERSITY OF
OXFORD

► Markov chain Monte Carlo (MCMC), a powerful collection of algorithms that enable us to simulate from complicated distributions using Markov chains.

► The development of MCMC has revolutionized statistics and scientific computation by vastly expanding the range of possible distributions that we can simulate from, including joint distributions in high dimensions.

► The basic idea is to build your own Markov chain so that the distribution of interest is the stationary distribution of the chain.

- We would like to sample from a distribution $p(\theta)$.
- Choose a symmetric proposal distribution q such that $q(x|y) = q(y|x)$ e.g. $q(y|x) = N(x, \sigma)$.
- choose a starting value for θ_0.

Metropolis Sampler

```
1       for t in 1 to T,
2           Sample θ* ~ q(.|θ^{t-1}) and u ~ Unif(0,1).
3           if u < p(θ*)/p(θ^{t-1}), then θ^t = θ* else θ^t = θ^{t-1}.
```

Use the Metropolis algorithm to sample from the posterior distribution of the example we looked at $p(\lambda|j_1, \ldots, j_m)$

$$p(\lambda|j_1, \ldots, j_m) = \left[\frac{b^a}{j_1! \ldots j_m! \Gamma(a) P(j_1, \ldots, j_m)} \right] \times$$
$$\lambda^{(a+j_1+\cdots+j_m)-1} e^{-(b+m)\lambda}$$

Go to R.

We usually don't know when the chain will reach its stationary distribution:

- ▶ Run multiple instance of the chain with different starting points and see if they converge to the same distribution.
- ▶ Remove the initial part of the chain (Burn-in)

Optimising certain criteria leads to the suggestion that the proposal distribution $q(.|x) = N(x, \sigma^2)$ has to be chosen such that

- ▶ acceptance rate is ≈ 0.5 for $d = 1, 2$.
- ▶ acceptance rate is ≈ 0.25 for $d \geq 3$.

To reduce the correlation between the samples we can use thinning which means using every nth sample and discarding the others.

A standard assumption when modelling genotypes of bi-allelic loci (e.g. loci with alleles A and a) is that the population is "randomly mating". From this assumption it follows that the population will be in "Hardy Weinberg Equilibrium" (HWE), which means that if p is the frequency of the allele A then the genotypes AA, Aa and aa will have frequencies $p^2, 2p(1 - p)$ and $(1 - p)^2$ respectively.

Suppose that we sample n individuals, and observe n_{AA} with genotype AA, n_{Aa} with genotype Aa and n_{aa} with genotype aa.

▶ Model the posterior distribution of p.
▶ Write MCMC routine to sample from the posterior distribution of p.

UNIVERSITY OF
OXFORD

► What is the likelihood function?
► What values can p take? What is a good prior for p?

- ▶ What is the likelihood function?
- ▶ What values can p take? What is a good prior for p?

Likelihood: $f(n_{AA}, n_{Aa}, n_{aa}|p) \propto p^{2n_{AA}} \times (2p(1-p))^{n_{Aa}} \times (1-p)^{2n_{aa}}$

Prior: $\qquad\qquad p \sim \text{Unif}(0,1)$

Posterior: $f(p|n_{AA}, n_{Aa}, n_{aa}) \propto p^{2n_{AA}} \times (2p(1-p))^{n_{Aa}} \times (1-p)^{2n_{aa}}$
$$\times 1$$

Go to R.

UNIVERSITY OF
OXFORD

A slightly more complex alternative than HWE is to assume that there is a tendency for people to mate with others who are slightly more closely-related than "random" (as might happen in a geographically-structured population, for example). This will result in an excess of homozygotes compared with HWE. A simple way to capture this is to introduce an extra parameter, the "inbreeding coefficient" g, and assume that the genotypes AA, Aa and aa have frequencies $gp + (1-g)p^2, (1-g)2p(1-p)$, and $g(1-p) + (1-g)(1-p)^2$.

Suppose that we sample n individuals, and observe n_{AA} with genotype AA, n_{Aa} with genotype Aa and n_{aa} with genotype aa.

▶ Model the joint posterior distribution of (p, g).
▶ Write MCMC routine to sample from the joint posterior distribution of (p, g).

Likelihood: $f(n_{AA}, n_{Aa}, n_{aa}|(p, g)) \propto (gp + (1-g)p^2)^{n_{AA}} \times$
$$((1-g)2p(1-p))^{n_{Aa}} \times$$
$$(g(1-p) + (1-g)(1-p)^2)^{n_{aa}}$$

Prior: Assume independence$(p, g) \sim \text{Unif}(0, 1) \times \text{Unif}(0, 1)$

Posterior:$f((p, g)|n_{AA}, n_{Aa}, n_{aa}) \propto (gp + (1-g)p^2)^{n_{AA}} \times$
$$((1-g)2p(1-p))^{n_{Aa}} \times$$
$$(g(1-p) + (1-g)(1-p)^2)^{n_{aa}}$$
$$\times 1 \times 1$$

Go to R.