Stat 499: Expectations and Sampling methods

Kai Gong

What is Sampling?

- the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population
- Keywords: **selection, population, estimation**

Importance Sampling

- **Approximate E[f] by drawing samples from a"proposal distribution" q**, and correcting appropriately by a weighting ratio.
- Suppose dealing with p(z) is harder, i.e., we **can't even evaluate p(z)** but can only do so up to proportionality constant, and only $\tilde{p}(z)$ can be evaluated. **We can still apply importance sampling by applying the importance weight**.

Rejection Sampling

- Need to setup **a proposal function q(z) and M**, so that **$Mq(z) \geq \tilde{p}(z)$**, for all z.
- Simulate **U~Unif(0,1)** and candidate X~g from the candidate density.
- Use **$U<\tilde{p}(z)/Mq(z)$** to test if reject the candidate X or not.

Bayesian inference

- A method of statistical inference in which **Bayes' theorem** is used to update the probability for a hypothesis as more evidence or information becomes available.
- **$P(\theta|D)=(P(D|\theta) \times P(\theta))/P(D)$**
- Here, **$P(\theta)$** is the **prior**, **$P(D|\theta)$** is the **likelihood** of observing our result given our distribution for θ. $P(D)$ is the evidence. **$P(\theta|D)$** is the **posterior belief** of our parameters after observing the evidence i.e the number of heads .
- Use **$P(\theta|D)$** to estimate the probability of θ given the data.

Maximum likelihood estimation

- A method of **estimating the parameters of a probability distribution by maximizing the likelihood function**, so that under the assumed statistical model the observed data is most probable.
- $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, \widehat{L}_n(\theta\,; \mathbf{y})$ where θ is the estimating parameter, y is the given data, and Ln is the likelihood function.
- In practice, it is often convenient to work with **log likelihood**.

Case study: Zero - Inflated Poisson (ZIP) sampling model

Kai Gong

Introduction to the problem

- In practice, we often observe **many zeros than we would predict** from a **Poisson model.**
- We will define a **ZIP model** and then find some unknown parameters of the ZIP model using a Bayesian approach.
- We will use the sampling methods (importance sampling and rejection sampling) to answer questions related to the parameters of interest

Introduction to dataset

- Given the dataset of fish count, we want to **model how many fish are being caught** by fishermen at a state park.
- Some visitors do not fish, but there is no data on whether a person fished or not.
- Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish.

Head of data

	nofish <int>	livebait <int>	camper <int>	persons <int>	child <int>	xb <dbl>	zg <dbl>	count <int>
1	1	0	0	1	0	-0.8963146	3.0504048	0
2	0	1	1	1	0	-0.5583450	1.7461489	0
3	0	1	0	1	0	-0.4017310	0.2799389	0
4	0	1	1	2	1	-0.9562981	-0.6015257	0
5	0	1	0	1	0	0.4368910	0.5277091	1
6	0	1	1	4	2	1.3944855	-0.7075348	0

6 rows

In this project, we only consider modeling the count variable. Here is the table and summary of the count variable.

0	1	2	3	4	5	6	7	8	9	10	11	13	14	15	16	21	22	29	30	31	32	38	65	149
142	31	20	12	6	10	4	3	2	2	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	3.296	2.000	149.000

Model expression

We are going to denote the count variable data as $y_1, \ldots, y_n$ Assume that $y_1, \ldots, y_n$ are independent and identically distributed given $0 < \pi < 1$ and $\lambda > 0$ according to a zero-inflated Poisson sampling model:

$$Pr(y_i = 0 | \pi, \lambda) = \pi + (1 - \pi) \exp(-\lambda)$$

$$Pr(y_i = k) = (1 - \pi) \frac{\lambda^k \exp(-\lambda)}{k!} \quad \text{for } k = 1, 2, \ldots$$

log likelihood function for the parameters

$$l(\pi, \lambda) = r_0 ln(\pi + (1 - \pi) e^{-\lambda}) + (1 - r_0) ln(1 - \pi) + \bar{y} ln(\lambda)$$

$$r_0 = \frac{1}{n} \sum_{i=1}^{n} II (y_i = 0)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

MLE for the likelihood function

Use the optim function in R:

$\boldsymbol{\pi}_{MLE}$ = 0.5677436

λ_{MLE} = 7.6252130

Notice that the λ calculated here is highly influenced by a outlier (149), thus it is biased. We want to

introduce Bayesian inference to eliminate the influence of outliers by adding prior to the model later.

Estimate the probability that a new fisher catches no fish using the MLE of the parameters:

$$Pr(y_i = 0 \mid \widehat{\pi}, \widehat{\lambda}) = \widehat{\pi} + (1 - \widehat{\pi})\, e^{-\widehat{\lambda}}$$

= 0.5679546

Assuming that the new fisher catches at least one fish, estimate the probability that a fisher catches exactly one fish:

$$Pr(y_i = k) = (1 - \widehat{\pi}) \frac{\lambda^k e^{-\widehat{\lambda}}}{k!} \quad for \quad k = 1, 2, \ldots$$

=0.003722854

Bayesian analysis

Suppose the prior distribution be: **π ~ Beta(α, β)**, **λ ~ Gamma(shape = θ, rate = κ)**, and assume independence. Assume that **$\alpha = \beta = \kappa = 1$ and $\theta = 4$**.

Then the log posterior of π, λ up to a proportionality constant shows as below:

$$ln(posterior(\pi, \lambda | y) = l(\pi, \lambda) + (\theta - 1)\, ln(\lambda) - \kappa\lambda$$

The MAP estimate:

$$\hat{\lambda} = 4.3798186 \quad \hat{\pi} = 0.5625373$$

Determine the posterior mean of π and λ using importance sampling

proposal distribution: π,λ ~multinormal(mean = 0.5,4 , sd matrix = (0.1,0;0,0.5))

Generate 10000 pair of data from the proposal distribution, calculate the qtilda=dmvnorm(param)

Then calculate the rtilda =posterior/qtilda= exp(logposterior-log(qtilda))

λ.mean = ∑(λ*rtilda)/∑rtilda

π.mean = ∑(π*rtilda)/∑rtilda

Rejection sampling

Use proposal function π~beta(2,2), and λ~gamma(8,2). Let π*λ be g(x) and exp(logposterior) be f(x). Use optim function to find M is around 10000.

Simulate 10000 candidates X~g from the candidate density.

Simulate U~Unif(0,1).

If $U \leq \frac{f(x)}{Mg(x)}$

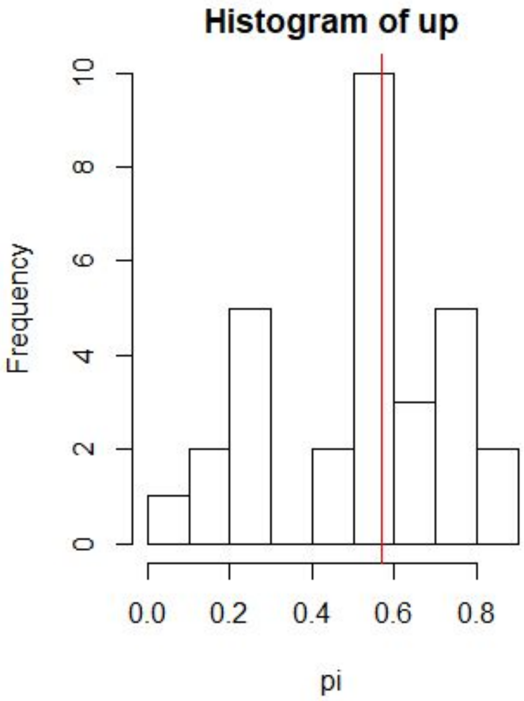
then “accept” the candidate X. Otherwise, “reject” X.

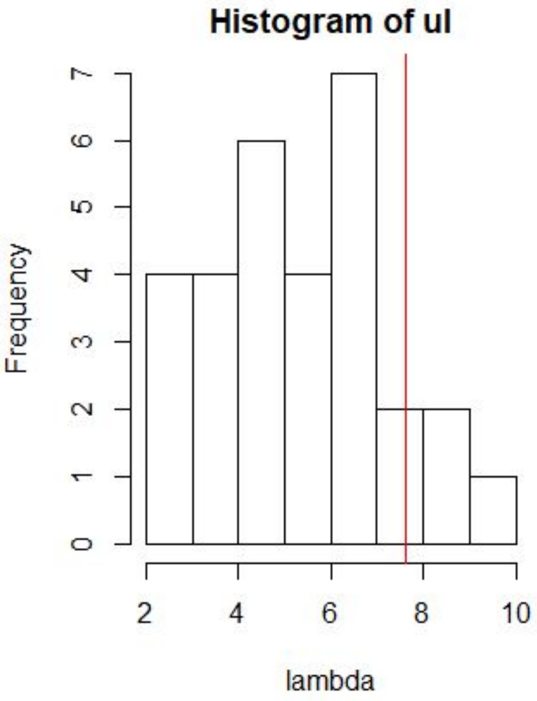
Comparison on statistics

	π	λ
MLE	0.5677436	7.6252130
MAP	0.5625373	4.3798186
Importance sampling	0.5242635	4.852539
Rejection sampling	0.5517844	4.875251

We can see that the result of using **importance sampling and rejection sampling does not show significant differences**. They **have an overlapped interval of CI for both π and λ**. The rejection sampling has a larger standard error because the proposal function can still be improved.

Visualization of samples using Rejection sampling





The red horizontal line indicates the MLE estimator for the parameters.

Using the posterior vs naive MLE

	Estimate the probability that a new fisher catches no fish using the MLE of the parameters	Assuming that the new fisher catches at least one fish, estimate the probability that a fisher catches exactly one fish
MLE parameter	0.5679546	0.003722854
Importance sampling	0.5355273	0.07013013
Rejection sampling	0.5227386	0.05840689

Summary

- In the case study, I used knowledge about Bayes Interface, Maximum likelihood, Importance sampling, and Rejection sampling.
- The result of case study show that these sampling techniques are useful in estimating the expectation of unknown distribution.
- Any questions?