HW_4_Maheshwari

Saurabh Maheshwari - 100%

May 24, 2018

Answer 1

All the calculations are done by hand, attached.

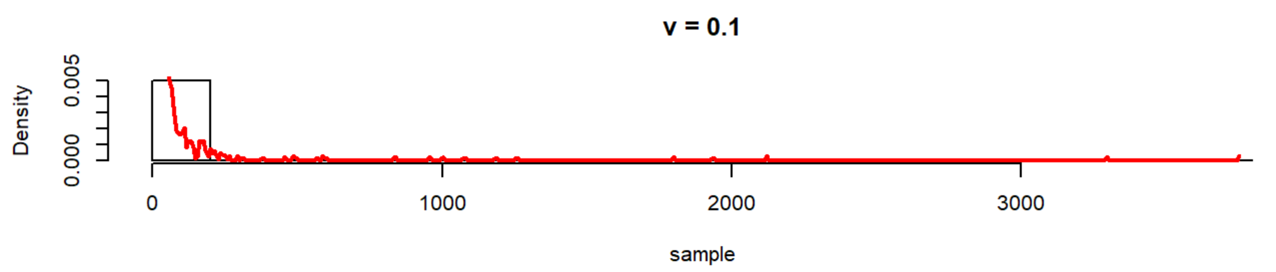
Answer 2

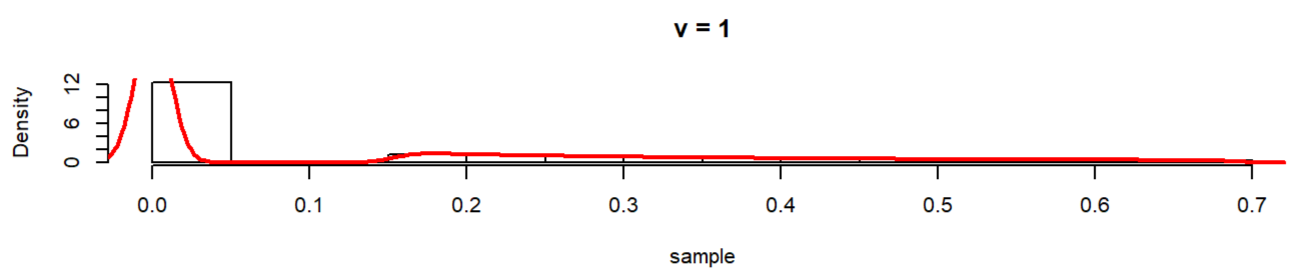
For each value of v, histograms of sampled values have been plotted. v = 0.1 and 10 had more extreme outliers as compared to v = 1 normally distributed g. For other details, have a look at the handwritten document attached. The code is provided.

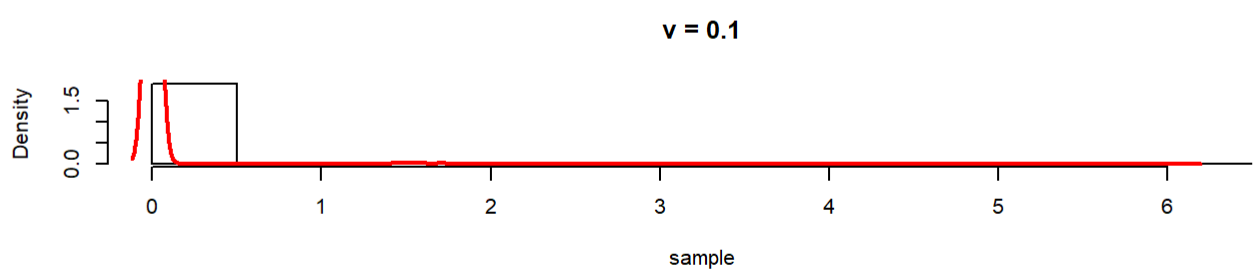
```
## [1] "Estimate when v = 0.1 =  0.137070639936199"
```

```
## [1] "Estimate when v = 1 =  0.135721673518058"
```

```
## [1] "Estimate when v = 10 =  0.135428759619688"
```







Answer 3

Part a) Monte Carlo Integration has been applied to estimate I_mc. The estimate comes out to be as follows. For more information have a look at the handwritten document attached.

```
## [1] "Monte carlo integration estimate, I_mc =   0.693504593478086"
```

part b, c) E(c(U)) has been calculated analytically and equals to 1.5. Control Variate estimate for I is given below. To calculate the variance, monte carlo and control variate estimates were calculated 1000 times, and variance of estimates has been calculated using these samples. It's observed that the variance of control variate estimate is much lesser than the monte carlo integration estimate. The code is provided.

```
## [1] "Control variate estimate for I =  0.69321383154079"
```

```
## [1] "Variance of Monte carlo integration =  1.27373066587352e-05"
```

```
## [1] "Variance of Control variate estimate =  2.08872717693301e-07"
```

part d) For an estimate with smaller variance, 2 approaches can be followed:

1. a larger sample could be used for calculating the estimate. In this problem, we have used n = 1500. Increasing n, could lead to estimate with smaller variance
2. Another control variate function showing better correlation with h(x) can be used. For example, c(x) = x/(x+1). In this case, the variance of control variate estimate is found to be 6.280189610^{-33}

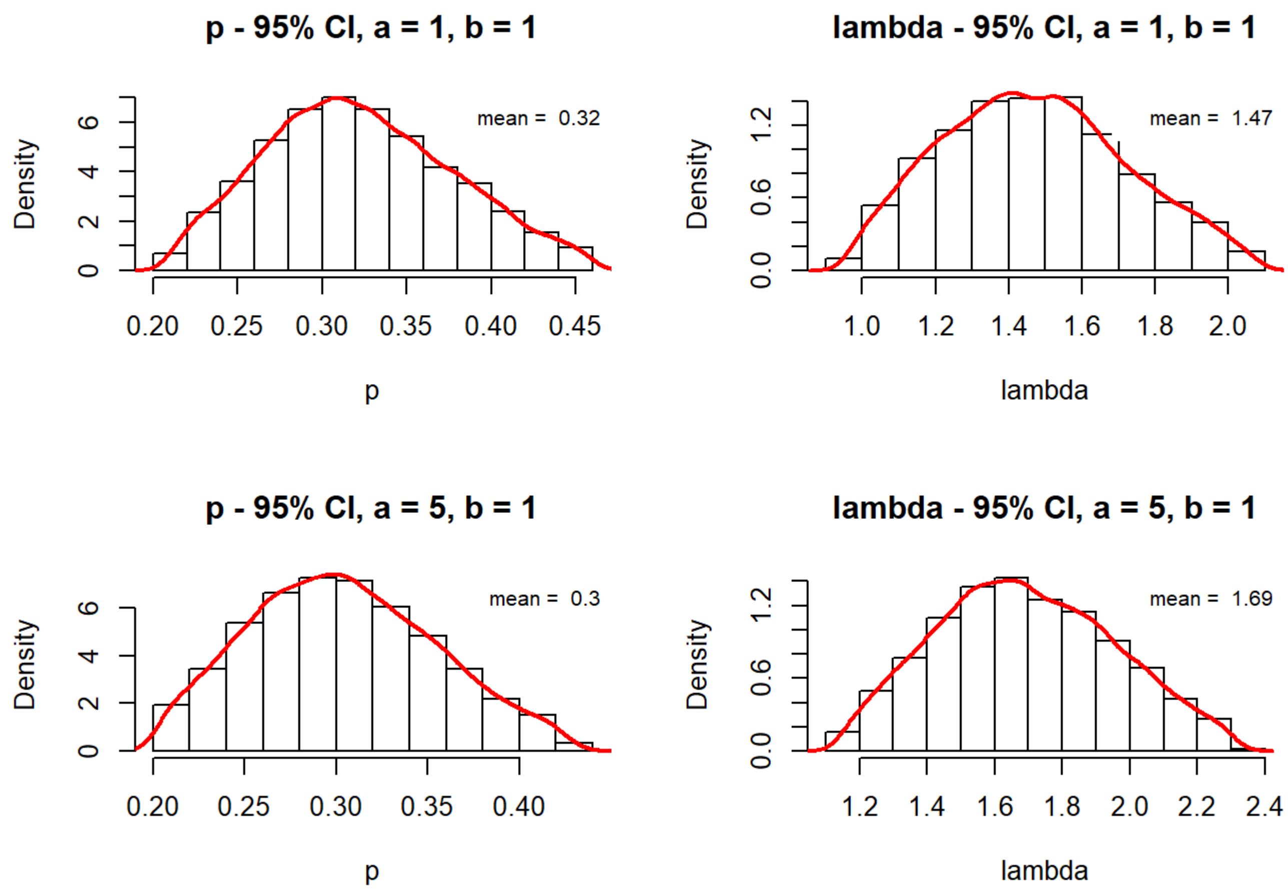
Answer 4

All the information is provided in the hand written document attached below.

Answer 5

part a) Random sample of size n = 100 is sampled using the ZIP model. First, r is calculated using bernoulli distribution with parameter p. Then, using the r_i, x_i was generated using poisson distribution with parameter $\lambda * r_i$. The code has been provided separately

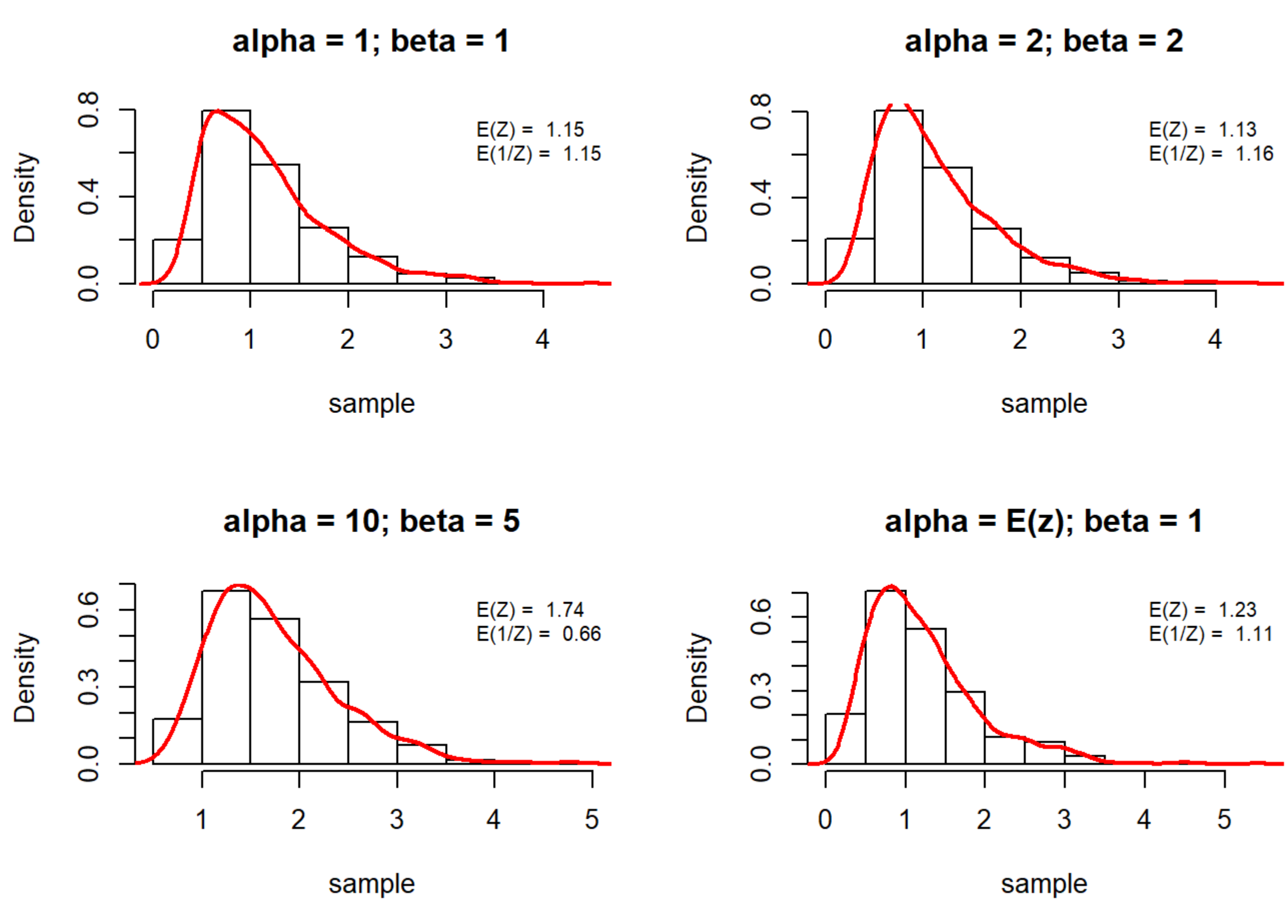
part b) All the proof have been provided in the hand written document part c) A sample of size 10000 was sampled using the gibbs sampler approach. 2 different kind of gamma distributions are used to generate λ. The sampled values from part a were used here as the observed data, and using it, λ , p and r_i are calculated. The histograms for λ and p after constructing 95% confidence intervals are plotted.



Thus, the real p and λ lie within the 95% confidence intervals drawn from the gibbs sampler. The code is provided

Answer 6

For kinds of gamma distributions are considered as the proposal density. The histogram for sample drawn from each distribution is plotted below. The comparison between the real values of E(z) and E(1/z) are mentioned in the plot below. The code is provided.



Comparing with the real E(z) = 1.154 and E(1/z) = 1.116, the gamma distribution with small shape parameter shows better results.

Ans. 1)

a) $I = \int_0^1 x^2 \, dx$

let $h(x) = x^2$; $f(x) = unif[0,1]$

Thus, $I = \frac{1}{n} \sum_{i=1}^{n} x_i^2$

Solving this in R with $n = 10{,}000$

$I = 0.3344$

b) $I = \int_0^1 \int_{-2}^{2} x^2 \cos(xy) \, dx \, dy.$

let $h(x) = x^2 \cos(xy)$; $f(x,y) = unif[-2,2] \times unif[0,1] = \frac{1}{(2+2)(1-0)} = \frac{1}{4}$

Thus, $I = \frac{1}{n} \sum_{i=1}^{n} x^2 \cos(xy_i) \times Area$; where $Area = 4 \times 1 = 4$

So, we sample $n = 10000$ x and y values using R, where ~~x is between~~ $x \sim unif[-2,2]$ and $y \sim [0,1]$.

$I = 3.4693$

c) $I = \int_0^\infty \frac{3}{4} x^4 e^{-x^3/3} \, dx$

$= \frac{3}{4} \Gamma(5) \int_0^\infty \frac{x^4 e^{-x}}{\Gamma(5)} \cdot e^{-x^3/3} \cdot e^{+x} \, dx$

Thus, $h(x) = e^{x - x^3/3}$ and $f(x) \sim gamma(5,1)$

$I = \Gamma(5) \frac{1}{n} \sum_{i=1}^{n} h(x_i) \times \frac{3}{4}$, $x \sim gamma(5,1)$

We, generate 10,000 x from gamma $(5,1)$ using R.

$I = \cancel{gamma} \; \Gamma(5) \times 0.75 \times \frac{1}{n} \sum_{i=1}^{n} h(x_i) = 1.3857$

Ans. 2) $I = \frac{1}{\sqrt{2\pi}} \int_1^2 e^{-x^2/2} dx$

$h(x) = e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}}$; $f(x) = \frac{1}{2-1} = 1$ (unif [0,1])

$g(x) = N(1.5, V^2)$

for, $V = 0.1$, $I = 0.1303$

$V = 1$, $I = 0.1356$

$V = 10$, $I = 0.1359$

In each case, $1e^6$ samples were drawn. The code in R has been provided.

Ans. 3)

a) $I = \int_0^1 \frac{1}{1+x} dx$

let $h(x) = \frac{1}{1+x}$, $f(x) = \frac{1}{1-0} = 1$; $x \sim$ unif [0,1]

Generating 1500 $x \sim$ Unif [0,1] ; $I = \frac{1}{n} \sum_{i=1}^{n} h(x_i) = 0.68934$

b) $\hat{I}_{cv} = \frac{1}{n} \sum_{i=1}^{n} h(U_i) - b\left[\frac{1}{n} \sum_{i=1}^{n} c(U_i) - E\{c(U)\}\right]$

$c(U) = 1+U \Rightarrow E(c(U)) = \int_0^1 (1+x) f(x) dx$, where $f(x) = \frac{1}{1-0} = 1$ as $x \sim$ Unif [0,1]

$= x + \frac{x^2}{2} \Big|_0^1 = \frac{3}{2}$

$$\hat{b} = \frac{cov(\hat{\mu}_{MC}, \hat{\theta}_{MC})}{var(\hat{\theta}_{MC})}$$

$$= \frac{\sum_{i=1}^{n} [h(\underset{\sim}{x_i}) - \hat{\mu}_{MC}][c(\underset{\sim}{y_i}) - \hat{\theta}_{MC}]}{\sum_{i=1}^{n} [c(\underset{\sim}{y_i}) - \hat{\theta}_{MC}]^2}$$

$$\left.\begin{aligned} \hat{\mu}_{MC} &= \frac{1}{n}\sum_{i=1}^{n} h(x_i) \\ \hat{\theta}_{MC} &= \frac{1}{n}\sum_{i=1}^{n} c(x_i) \end{aligned}\right\} x_i \sim unif[0,1]$$

Solving in R; $\hat{b} = -0.47864$

$\hat{I}_{cv} = 0.69395$

c) To compute variance, $\hat{I}_{MC}$ and $\hat{I}_{cv}$ were ~~s~~calculated 1000 times.

variance of $\hat{I}_{MC} = 1.29 \times 10^{-5}$

variance of $\hat{I}_{cv} = 3.95 \times 10^{-7}$

clearly, $Var(\hat{I}_{cv}) < var(\hat{I}_{MC})$

d) There are 2 ways of doing so,

1) increase the number of sample, i.e, n.

2) choose a control variate that has a better correlation than the current $c(x)$, for.eg; $c(x) = x/x+1$. In this case the variance comes out to be ~~around~~ around $1e^{-33}$

Answer 4)

a) F-test can not be used in this case because of non-normality of error terms. Thus, another statistic known as log likelihood statistics ~~is~~ can be used.

H_0 = Null hypothesis : $\alpha_1 = \alpha_2 = \alpha_3$; i.e, all the α's are same

H_A = Alternate hypothesis : Not H_0.

$$\Lambda = \frac{\sup_{\theta \in \Omega_0} \left(\frac{1}{2\sigma}\right)^n \exp\left[-\sum_{i=1}^{3}\sum_{j=1}^{n_i} \frac{|y_{ij}-\mu|}{\sigma}\right]}{\sup_{\theta \in \Omega} \left(\frac{1}{2\sigma}\right)^n \exp\left[-\sum_{i=1}^{3}\sum_{j=1}^{n_i} \frac{|y_{ij}-\mu_i|}{\sigma}\right]}$$

$$= \left(\frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i} |y_{ij} - med_i|}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} |y_{ij} - med|}\right)^n$$; where med_i is the median of each group. and med is the median of all y_{ij}

Thus, if H_0 is true:

$$\Lambda = \left(\frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i} |e_{ij} - med_i^e|}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} |e_{ij} - med^e|}\right)^n$$; where med_i^e is the median of e_{ij} for group i and med^e is the median of all e_{ij}

Thus, for MC testing follow the steps:

1) Calculate $e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, e_{31}, \ldots, e_{3n_3}$
2) Calculate Λ assuming H_0 to be true.
3) Repeat steps to collect a large sample of Λ
4) If Λ (observed) lies amongst the $100(1-\alpha)\%$ of the sampled Λ, then reject H_0 at α significance level.

b) Randomization tests can be used without making any assumptions about the distribution of error terms.

$$\text{Test statistic} = \sum_{i>j} |\mu_i - \mu_j| \quad (i,j = 1,2,3)$$

H_0: Test statistic $= 0$

H_A: Not H_0

Steps:-

1) Combine all the observations to form a sample of size $n_1+n_2+n_3$
2) Randomly select n_1 observations without replacement, then n_2 observations to form 2^{nd} group and the remaining group will be the 3^{rd} group.
3) Calculate $\sum_{i>j} |\mu_i - \mu_j|$; $i,j \in (1,2,3)$. In case H_0 is true, these values should be close to zero.
4) Repeat 2, 3 steps to form a sample of test statistics
5) If 0 is not in the $100(1-\alpha)\%$ confidence interval, reject H_0 at α significance level.

Ans.5)

a) $p = 0.3$, $\lambda = 2$

First we calculate r_i by generating 100 bernoulli observations with probability of success $= p$.

then, we calculate x_i by generating poisson distribution sample with parameter $\lambda \times r_i$

b)(i) $$f(x, r, \lambda, p) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)} \prod_{i=1}^{n} \frac{e^{-\lambda r_i} (\lambda r_i)^{x_i}}{x_i!} p^{r_i} (1-p)^{1-r_i}$$

considering p, r, x as constants.

$$f(\lambda \mid p, r, x) = \frac{b^a \lambda^{(a+\sum_i x_i)-1} e^{-\lambda(b+\sum_i r_i)}}{(b+\sum_i r_i)^{(a+\sum x_i)} \Gamma(a+\sum x_i) / (b+\sum_i r_i)^{(a+\sum x_i)} \Gamma a} \cdot \Gamma(a+\sum x_i) \cdot \prod p^{r_i}(1-p)^{1-r_i}$$

$(\lambda \mid p, r, x) \sim \text{Gamma}(a+\sum_i x_i, b+\sum_i r_i)$ [as everything else is constant]

(ii) Considering λ, r, x as constants

$$f(p \mid \lambda, r, x) = \left[\frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)} \prod_{i=1}^{n} \frac{e^{-\lambda r_i}(\lambda r_i)^{x_i}}{x_i!}\right] \cdot \frac{p^{\sum r_i} \cdot (1-p)^{n-\sum r_i}}{B(\sum r_i+1, n-\sum r_i+1)} \, B(\sum r_i+1, n-\sum r_i+1)$$

$(p \mid \lambda, r, x) \sim \text{beta}(\sum r_i+1, n-\sum r_i+1)$

(iii) considering λ, p, x as constant

$$f(r_i \mid \lambda, p, x) = \text{constant} \times \prod_{i=1}^{n} \frac{e^{-\lambda r_i}(\lambda r_i)^{x_i}}{x_i!} p^{r_i}(1-p)^{1-r_i}$$

Since, $r_i \sim$ Bernoulli, it belongs to $\{0 \text{ or } 1\}$

Consider 2 cases:

I) If $x_i = 0$

$$\Rightarrow f(r_i \mid \lambda, p, x) = \text{constant} \times \prod_{i=1}^{n} e^{-\lambda r_i} \cdot p^{r_i} (1-p)^{1-r_i}$$

$$= \text{constant} \times \prod_{i=1}^{n} \left(\frac{e^{-\lambda} \cdot p}{e^{-\lambda} p + (1-p)} \right)^{r_i} \cdot \left(\frac{1-p}{e^{-\lambda} p + (1-p)} \right)^{1-r_i}$$

$$\Rightarrow (r_i \mid \lambda, p, x) \sim \text{Bernoulli} \left(\frac{e^{-\lambda} p}{e^{-\lambda} p + (1-p} \right)$$

II) If $x_i \neq 0$ and $r_i = 0$

$$\Rightarrow f(r_i \mid \lambda, p, x) = 0 \rightarrow \text{Not Possible.}$$

Thus, if $x_i \neq 0$ then $r_i = 1$ with probability $= 1$

Hence, $$(r_i \mid \lambda, p, x) \sim \text{Bernoulli} \left(\frac{e^{-\lambda} p}{e^{-\lambda} p + (1-p) \cdot I\{x=0\}} \right)$$

c) A sample of size $n = 10000$ is sampled. To construct 95% confidence interval 2.5% of data from each end were removed. The histograms have been plotted for 2 different pairs of a and b.

The original a and b lie within the confidence intervals.

Ans. 6) To draw independent-Metropolis-Hasting samples, 4 different gamma distributions are considered. The summary for each distribution is provided alongside the histograms.

Distribution 1 $\sim$ gamma(1,1)

" 2 $\sim$ gamma(2,2)

" 3 $\sim$ gamma(10,5)

" 4 $\sim$ gamma($\sqrt{t_2/t_1}$, 1)