THE UNIVERSITY OF NEWCASTLE AUSTRALIA

FACULTY OF SCIENCE AND INFORMATION TECHNOLOGY
School of Mathematical and Physical Sciences

Student Enquiries
Telephone: (02) 4921-5000

STAT3120 APPLIED BAYESIAN METHODS

Lab Exercises 6 Solutions

Q1 *The table below is taken from BDA. It contains results of a survey on bicycle traffic around Berkeley, CA in 1993. From the table in BDA we consider only the first four rows of data for residential streets. This consists of bicycle counts on a particular day on different streets around Berkeley. Also recorded is whether the street has a certified bicycle lane or not.*

Bike route?	Bicycle count	Bike route?	Bicycle count
Yes	16	Yes	55
Yes	9	No	12
Yes	10	No	1
Yes	13	No	2
Yes	19	No	4
Yes	20	No	9
Yes	18	No	7
Yes	17	No	9
Yes	35	No	8

We wish to compare the amount of bicycle traffic on the sections of road that are designated bike routes, with that for roads without bike routes.

Type the data in R:

```
y<-c(16,9,10,13,19,20,18,17,35,55)  # for bike route
z<-c(12,1,2,4,9,7,9,8)  # for non-bike route
```

(a) Consider the group bike route =’Yes’ (Y) first. What distribution are the bicycle counts most likely to follow?

Assuming the counts are independent and identically distributed, we would use a Poisson distribution here. So, Y ~ Poisson (θ_Y)

(b) Assume that the bicycle counts on bike routes are independent and identically distributed for different streets. What is the conditional likelihood formula for the counts Y?

$$p(y \mid \theta_Y) = \prod \frac{\theta_Y^{y_i} \exp(-\theta_Y)}{y_i!} = \frac{\theta_Y^{n\bar{y}} \exp(-n\theta_Y)}{\prod y_i!} \propto \theta_Y^{212} \exp(-10\theta_Y)$$

This is in the form of a Gamma(213,10) density in θ_Y.

(c) What prior information do we have? Construct a prior distribution for θ_Y the parameter of interest for the counts *y*.

In this case we seem to have no prior information about θ_Y. Our preferred noninformative prior is Gamma(1,0).

(d) Update the prior with the observed counts to find the posterior distribution for $\theta_Y \mid y$. Generate an estimate and 95% interval for $\theta_Y \mid y$.

The posterior for $\theta_Y \mid y$ is Gamma (213, 10). E($\theta_Y \mid y$) = 213/10 = 21.3

```
qgamma(0.025,213,10) = 18.54 ;  qgamma(0.975,213,10) = 24.25
```

A 95% interval is (18.54, 24.25)

(e) Let Z = bicycle counts on non-bike routes. How can we compare Y and Z?

Naturally, we compare θ_Y and θ_Z, the rates of bicycle usage for bike routes and non-bike routes.

(f) Estimate a 95% confidence interval on the difference $\theta_Y - \theta_Z$. What conclusions can you draw based on this interval?

Rather than attempt to find the true distribution of $\theta_Y - \theta_Z \mid y,z$ we can use Monte Carlo simulation here. Just simulate from $\theta_Y \mid y$ ~Gamma(213,10) and
$\theta_Z \mid z$ ~ Gamma (53, 8). Then simply subtract each iterate of θ_Z from each iterate of θ_Y obtaining a sample from $\theta_Y - \theta_Z \mid y,z$.

i.e.
```
ty<-rgamma(10000,213,10)
tz<-rgamma(10000,53,8)
tymz<-ty-tz
```

The quantile function is then used as follows

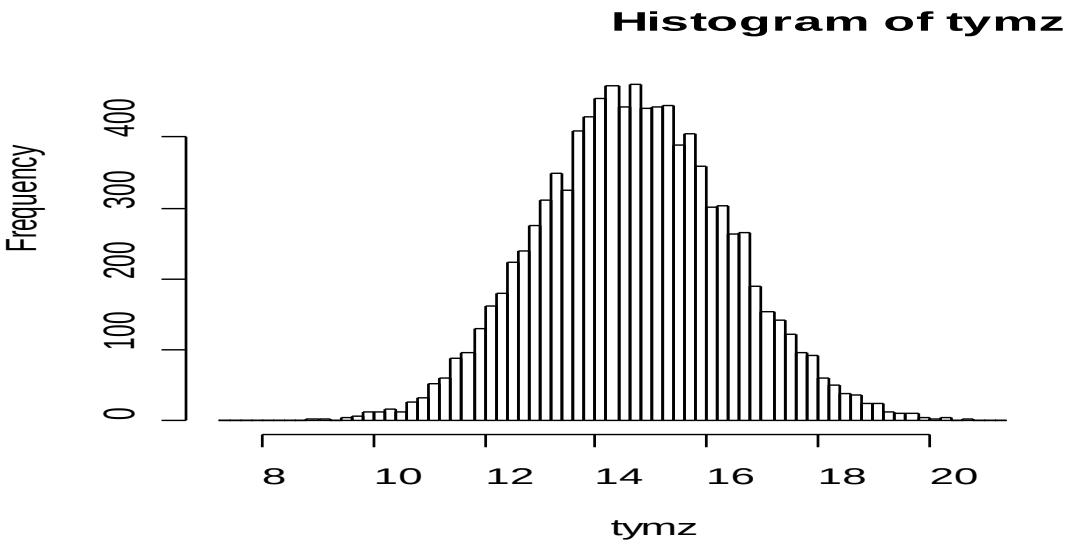
```
quantile(tymz,0.025) = 11.404
quantile(tymz,0.975) = 18.110
```

A point estimate is

```
mean(tymz) = 14.69
```

A histogram follows

```
hist(tymz)
```



Clearly bike routes are used significantly more than non-bike routes.

(g) Estimate the probability that $\theta_Y - \theta_Z$ is greater than 10 given the data.

Using the sample in (f)
`length(tymz[tymz>10])/length(tymz) = 0.997` (or: mean(tymz>10)!)

We are very confident that bike usage is at least 10 per day per street more on bike routes than non-bike routes.

(h) Perform the equivalent t-test. Discuss why inference appears different from (g).

`t.test(y,z,"greater",10)` gives a p-value of 0.1645. This would need to be compared with 0.003 based on (g)! In hindsight, there is a problem with the Poisson assumption (rather than the t-test, in our opinion, as it's very robust). E.g. mean(y) = 21.2 and var(y) = 192.8: there seems to be overdispersion.

Q2 (a) The likelihood contributes a Gamma (213, 10) to the posterior.

A Gamma(a,b) prior means the posterior is $\theta_Y \mid y \sim$ Gamma (212+a, 10+b)

So we have Gamma(0,0) prior → $\theta_Y \mid y \sim$ Gamma (212, 10)

And Gamma(0.5,0) prior → $\theta_Y \mid y \sim$ Gamma (212.5, 10)

95% intervals for these are (using m=10,000,000 – even then not always enough for 2 dps!)

prior	posterior	95% interval $\theta_Y \mid y$	95% interval $\theta_Y - \theta_Z \mid y,z$	Pr($\theta_Y - \theta_Z$ >10)
0,0	212,10	18.44, 24.15	11.38, 18.09	0.997
0.5,0	212.5,10	18.49, 24.20	11.36, 18.09	0.997
1,0	213,10	18.54, 24.25	11.33, 18.08	0.997

Results seem quite insensitive to these prior choices.

R code

```
## For Gamma(0,0)
qgamma(c(0.025,0.975),212,10) ##


ty1=rgamma(10000,212,10)
tz1=rgamma(10000,52,8)
tymz1=ty1-tz1

## The quantile function is then used as follows
quantile(tymz1,c(0.025,0.975)) ##

length(tymz1[tymz1>10])/length(tymz1)  ###

## For Gamma(0.5,0)
qgamma(c(0.025,0.975),212.5,10) ##

ty2=rgamma(10000,212.5,10)
tz2=rgamma(10000,52.5,8)
tymz2=ty2-tz2

## The quantile function is then used as follows
quantile(tymz2, c(0.025,0.975)) ##

length(tymz2[tymz2>10])/length(tymz2)  ###
```

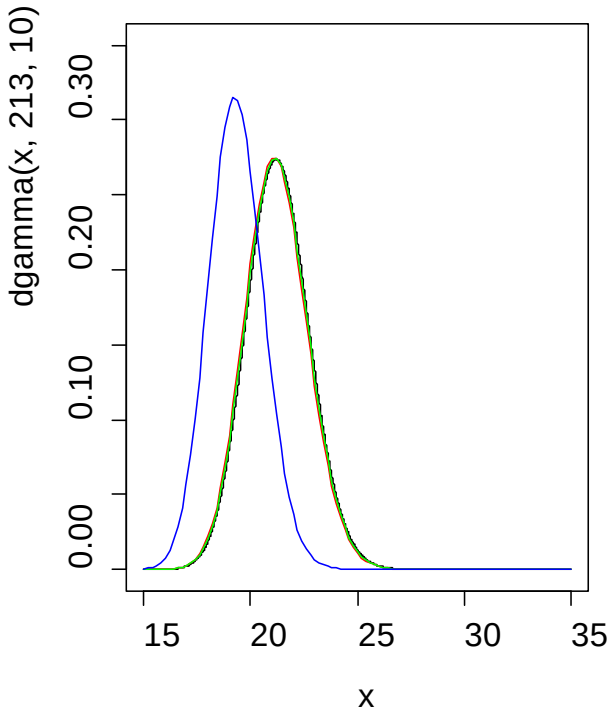
(b) Under a Gamma(20,2) prior $\theta_Y \mid y \sim$ Gamma (232, 12) and under a Gamma(20,3) prior $\theta_Z \mid z \sim$ Gamma (72, 11)

```
## For informative priors Gamma(20,2) for Y and Gamma(20,3) for Z
## For Y
qgamma(c(0.025,0.975),232,12) ##

ty3=rgamma(10000,232,12)
tz3=rgamma(10000,72,11)
tymz3=ty3-tz3

## The quantile function is then used as follows
quantile(tymz3, c(0.025,0.975)) ##

length(tymz3[tymz3>10])/length(tymz3)  ###
```

The table now is

Prior	posterior	95% interval $\theta_Y \mid y$	95% interval $\theta_Y - \theta_Z \mid y, z$	$\Pr(\theta_Y - \theta_Z > 10)$
0,0	212,10	18.44, 24.15	11.38, 18.09	0.997
0.5,0	212.5,10	18.49, 24.20	11.36, 18.09	0.997
1,0	213,10	18.54, 24.25	11.33, 18.08	0.997
20,2 20,3	232, 12 72, 11	16.93, 21.90	9.91, 15.74	0.971

Results have changed somewhat with the expert prior. The prior on θ_Y has a mean at 20/2 = 10 which is well below the mean of y (21.3), which explains why the posterior has been shrunk towards 0.



The posteriors for $\theta_Y \mid y$ are shown above. The blue line is the Gamma(232,12) from the expert prior. The posteriors based on the 3 candidate noninformative priors are almost indistinguishable.

```
x=seq(15,35,0.01)
plot(x, dgamma(x,213,10),type="l",ylim=c(0,0.35))
curve(dgamma(x,212,10),col="red", add=TRUE)
curve(dgamma(x,212.5,10),col="green", add=TRUE)
curve(dgamma(x,232,12),col="blue", add=TRUE)
```