RESTRICTED OPEN BOOK EXAMINATION
Data provided: Table of Normal Quantiles

MAS364

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2013–14**

Bayesian Statistics **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
Marks will be awarded for your best ***three*** *answers. Total marks 84.*

Standard results from the lecture notes may be used without derivation, but must be clearly stated.

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 (i) A zoologist wants to learn about the true weight in grammes, θ, of an animal of a newly discovered species. Based on a visual assessment and knowledge of related species, his best guess for θ is 1000, and he thinks that it is 95% probable that θ lies between 900 and 1100.

(a) Find a suitable normal distribution to represent this prior distribution for θ. ***(3 marks)***

(b) He then takes a measurement X of the weight of the animal in grammes, using equipment known to have errors with mean zero and standard deviation 40, so that $X \sim N(\theta, 40^2)$. State the zoologist's posterior distribution for θ after observing $X = x$. ***(3 marks)***

(c) If $x = 920$, calculate his posterior mean and variance for θ. Find values z_L and z_U such that his posterior probabilities for $\theta < z_L$ and $\theta > z_U$ are both equal to 0.25. How do these values compare with the corresponding quantiles of his prior distribution? ***(8 marks)***

(ii) A unitless physical constant ψ is well known from a combination of different experimental results; experts are agreed on a prior distribution which is normal with mean $\mu = 0.23120$ and standard deviation $\tau = 1.5 \times 10^{-4}$.

(a) If an estimate $\hat{\psi}$ is to be made from this prior, based on a quadratic loss function, give the optimal value of the estimate and the expected loss incurred. ***(4 marks)***

(b) If observations $X_1, \ldots, X_n$ are to be made, with $X_i \sim N(\psi, \sigma^2)$ where $\sigma = 2 \times 10^{-3}$ (and the observations are conditionally independent given ψ), how large must n be to allow an estimate with *half* the quadratic loss of that based on the prior? ***(6 marks)***

(c) What is the predictive distribution for the first measurement, X_1, based on the prior information only? ***(4 marks)***

2 A horticulturalist is interested in the probability θ that a seed of a particular variety germinates successfully. Her prior distribution for the germination probability can be represented by the Beta(a, b) distribution; in an experiment she then observes n (conditionally independent) seeds, of which x germinate successfully.

(i) Write down her posterior distribution for θ. ***(2 marks)***

(ii) If her prior is determined by $a = 3, b = 1$, and she observes $x = 7$ successes with $n = 10$ seeds, give her posterior distribution and posterior mean and variance for θ. ***(4 marks)***

(iii) What is her predictive probability that the next seed observed, after the experiment above, germinates sucessfully? What is her predictive probability that *all* the seeds in a further batch of 10 would germinate? What would the corresponding probabilities have been based only on her prior beliefs? ***(8 marks)***

(iv) She wishes to set up a further experiment based on m seeds. Show that her probability for one or more seeds germinating is

$$1 - \frac{(m+3)\times(m+2)\times\cdots\times 4}{(m+13)\times(m+2)\times\cdots\times 14}$$

and hence show that she would require $m \geq 5$ to ensure that the probability of one or more seeds germinating is at least 0.99. ***(8 marks)***

(v) A less experienced colleague wants to repeat the above analyses but has little knowledge of germination rates for seeds of this sort. Explain how you would modify your analysis above in this case, and what differences you would expect in the numerical results in (ii) and (iii), and the required value of m in (iv); no further calculation is required. ***(6 marks)***

3 The operational lifetime in hours, Y, of a particular kind of drill is known to follow an exponential distribution with rate parameter θ, so that

$$f(y|\theta) = \theta \exp(-y\theta), \quad y \geq 0$$

and

$$F(y|\theta) = 1 - \exp(-y\theta), \quad y \geq 0.$$

A manufacturing process can be carried out either manually (decision d_0) which has utility 10, or using a drill of the above type (decision d_1) in which case the utility is 20 if the drill operates for long enough (if $Y \geq 8$) or -15 if the drill fails during the process (if $Y < 8$).

(i) If θ is *known*, show that d_1 is optimal if and only if θ is below some threshold θ_0, and obtain the numerical value of θ_0. ***(7 marks)***

(ii) In practice, θ is *unknown*, with prior distribution Gamma(a, b), so that

$$f(\theta) = \frac{b^a \theta^{a-1} \exp(-b\theta)}{\Gamma(a)}, \quad \theta > 0.$$

If observations $X_1, \ldots, X_n$ are made, each Exponential(θ) and conditionally independent of each other, and of Y, given θ, derive the form of the posterior distribution for θ given $X_1, \ldots, X_n$. ***(5 marks)***

(iii) Past experience with other kinds of drill suggests that the parameter θ is most likely to be around 0.07, with standard deviation 0.025; three observations on the current type give lifetimes of 27 hours, 21 hours and 16 hours. Write down the prior and posterior distributions for θ given this information. ***(8 marks)***

(iv) For the general case of observations $X_1, \ldots, X_n$, derive the predictive distribution for Y. ***(6 marks)***

Without further calculation, explain how the decision in (i) should be made in the light of uncertainty about θ. ***(2 marks)***

4 The table below shows data on the numbers of work-related accidents $A_1, \ldots, A_8$ occurring within a sample of similar-sized companies in the same industry, recorded over a year, as part of a study about accident rates in the industry as a whole, which is made up of a much larger number of companies.

Index i	1	2	3	4	5	6	7	8
A_i	54	19	44	60	49	51	20	70

The WinBUGS code below implements a model intended to help with the interpretation of these data.

```
model
{
for (j in 1:N)
{
L[j]~dnorm(M,P)
R[j]<-exp(L[j])
A[j]~dpois(R[j])
}
M~dnorm(3,0.25)
P~dgamma(1,4)
V<-1/P
S<-sqrt(V)
}
```

The model is to be run using the following data:

```
list(N=8,A=c(54,19,44,60,49,51,20,70))
```

(i) Write down the model in mathematical terms, and draw a directed acyclic graph to represent its structure. ***(10 marks)***

(ii) A simpler model could be expressed in WinBUGS as follows.

```
model
{
for (j in 1:N)
{
L[j]~dnorm(0,0.001)
R[j]<-exp(L[j])
A[j]~dpois(R[j])
}
}
```

Explain briefly the key statistical differences between the models and their implications for this analysis. ***(5 marks)***

 Question 4 continued on next page

4 (continued)

(iii) The table below shows statistical summaries (in WinBUGS) of some of the output from running the model.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
L[1]	3.974	0.1363	0.001321	3.699	3.975	4.229	1001	10000
L[2]	2.96	0.2265	0.002139	2.495	2.967	3.376	1001	10000
L[3]	3.773	0.1504	0.001427	3.465	3.776	4.054	1001	10000
L[4]	4.08	0.1306	0.001223	3.815	4.083	4.327	1001	10000
L[5]	3.878	0.1439	0.001263	3.588	3.882	4.157	1001	10000
L[6]	3.919	0.1392	0.001389	3.637	3.923	4.188	1001	10000
L[7]	3.005	0.2189	0.00203	2.553	3.013	3.414	1001	10000
L[8]	4.232	0.1198	0.001179	3.991	4.233	4.463	1001	10000
M	3.705	0.4147	0.004599	2.831	3.708	4.524	1001	10000
R[1]	53.69	7.289	0.07072	40.42	53.25	68.67	1001	10000
R[2]	19.79	4.42	0.04223	12.13	19.44	29.26	1001	10000
R[3]	43.99	6.578	0.06294	31.97	43.66	57.65	1001	10000
R[4]	59.66	7.763	0.07137	45.39	59.34	75.75	1001	10000
R[5]	48.85	6.995	0.06093	36.16	48.51	63.88	1001	10000
R[6]	50.85	7.048	0.07051	37.96	50.54	65.91	1001	10000
R[7]	20.66	4.465	0.04041	12.85	20.36	30.39	1001	10000
R[8]	69.33	8.275	0.08152	54.12	68.94	86.7	1001	10000
S	1.135	0.3796	0.005883	0.5375	1.086	1.993	1001	10000

(a) Based on the table, give 95% central posterior intervals for the annual accident rate for company 1, and for the underlying average accident rate for the industry. *(5 marks)*

(b) What can you say about the variability of accident rates between companies? *(4 marks)*

(c) How do the prior and posterior distributions for the quantity `M` compare? Explain whether you think the prior distribution for `M` seems reasonable. *(4 marks)*

End of Question Paper

Table of quantiles of the standard normal distribution

q	$\Phi^{-1}(q)$	q	$\Phi^{-1}(q)$	q	$\Phi^{-1}(q)$
0.0000	$-\infty$	0.3375	-0.419	0.6750	0.454
0.0125	-2.241	0.3500	-0.385	0.6875	0.489
0.0250	-1.960	0.3625	-0.352	0.7000	0.524
0.0375	-1.780	0.3750	-0.319	0.7125	0.561
0.0500	-1.645	0.3875	-0.286	0.7250	0.598
0.0625	-1.534	0.4000	-0.253	0.7375	0.636
0.0750	-1.440	0.4125	-0.221	0.7500	0.674
0.0875	-1.356	0.4250	-0.189	0.7625	0.714
0.1000	-1.282	0.4375	-0.157	0.7750	0.755
0.1125	-1.213	0.4500	-0.126	0.7875	0.798
0.1250	-1.150	0.4625	-0.094	0.8000	0.842
0.1375	-1.092	0.4750	-0.063	0.8125	0.887
0.1500	-1.036	0.4875	-0.031	0.8250	0.935
0.1625	-0.984	0.5000	0.000	0.8375	0.984
0.1750	-0.935	0.5125	0.031	0.8500	1.036
0.1875	-0.887	0.5250	0.063	0.8625	1.092
0.2000	-0.842	0.5375	0.094	0.8750	1.150
0.2125	-0.798	0.5500	0.126	0.8875	1.213
0.2250	-0.755	0.5625	0.157	0.9000	1.282
0.2375	-0.714	0.5750	0.189	0.9125	1.356
0.2500	-0.674	0.5875	0.221	0.9250	1.440
0.2625	-0.636	0.6000	0.253	0.9375	1.534
0.2750	-0.598	0.6125	0.286	0.9500	1.645
0.2875	-0.561	0.6250	0.319	0.9625	1.780
0.3000	-0.524	0.6375	0.352	0.9750	1.960
0.3125	-0.489	0.6500	0.385	0.9875	2.241
0.3250	-0.454	0.6625	0.419	1.0000	∞