Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2013–2014**

MAS6004 Inference **3 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 **five** answers. Total marks 100.*

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. ***(2 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 $X = x$. ***(2 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? ***(6 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. ***(3 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? ***(4 marks)***

(c) What is the predictive distribution for the first measurement, X_1, based on the prior information only? ***(3 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. ***(1 mark)***

(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. ***(3 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? ***(5 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

$$1 - \frac{(m+3)\times(m+2)\times\cdots\times 4}{(m+13)\times(m+12)\times\cdots\times 14}$$

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. ***(6 marks)***

(v) A less experienced colleague wants to repeat the above analyses but has little knowledge of germination rates for seeds of this sort. Give two different distributions that would be suitable for representing this prior ignorance, and comment briefly on their advantages and disadvantages. Explain what differences you would expect this change to make to the numerical results in (ii); no further calculation is required. ***(5 marks)***

3 (i) 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))
```

Write down the model in mathematical terms, and draw a directed acyclic graph to represent its structure. ***(7 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 the analysis of the data in (i). ***(3 marks)***

3 (continued)

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

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. ***(4 marks)***

(b) What can you say about the variability of accident rates across the industry? ***(3 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. ***(3 marks)***

4 (i) Suppose it is desired to sample from the Student's t-distribution with 7 degrees of freedom. The density is given by

$$f(x) = \frac{16}{5\sqrt{7}\pi}\left(1 + \frac{x^2}{7}\right)^{-4}.$$

Importance sampling is to be used, with an importance density based on approximating $f(x)$ by a normal density function.

(a) By considering a Taylor series expansion of $\log f(x)$ about 0, obtain the mean and variance of the importance density. ***(5 marks)***

(b) Given a random draw $Z_1 = -0.52$ from a $N(0,1)$ distribution, obtain a samples from the t_7 distribution via the importance density found above, and calculate the weight of your sampled value. ***(4 marks)***

(ii) Wind speeds (in m/s) are measured at a location at noon over the course of a week and observed to be $\{10.90, 26.04, 15.18, 15.46, 7.45, 15.76, 8.90\}$. A Weibull density is fitted to these data:

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The log-likelihood is maximised at $k = 2.65, \lambda = 16.07$. By considering the profile deviance function, test the null hypothesis that $k = 2$.

$\Big[$Note that: $\sum_i \log(x_i) = 18.06$, $\sum_i x_i^2 = 1649$ and $\sum_i x_i^{2.65} = 10995$. $\Big]$

(11 marks)

5 (i) Let Y be an exponential random variable with rate λ and density function

$$f_Y(y) = \begin{cases} \lambda \exp(-\lambda y) & \text{for } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Explain how to generate a random value of Y given a uniform random number using the inversion method. ***(4 marks)***

(ii) The failure time $t_1, \ldots, t_4$ of four identical light bulbs are claimed to be independent and follow an exponential distribution with rate 4. We wish to test the assumption of independence as the failure times seem very similar.

(a) Assuming the model of an Exponential distribution is correct, explain carefully how a Monte-Carlo test of size 0.05 could be constructed to test the hypothesis of independence, using the sample variance as a test statistic. ***(3 marks)***

(b) Given the uniform random numbers $0.12, 0.61, 0.63, 0.11$ generate one random value of the test statistic under H_0. ***(4 marks)***

(iii) The $Gamma(5, 2)$ density function is given by

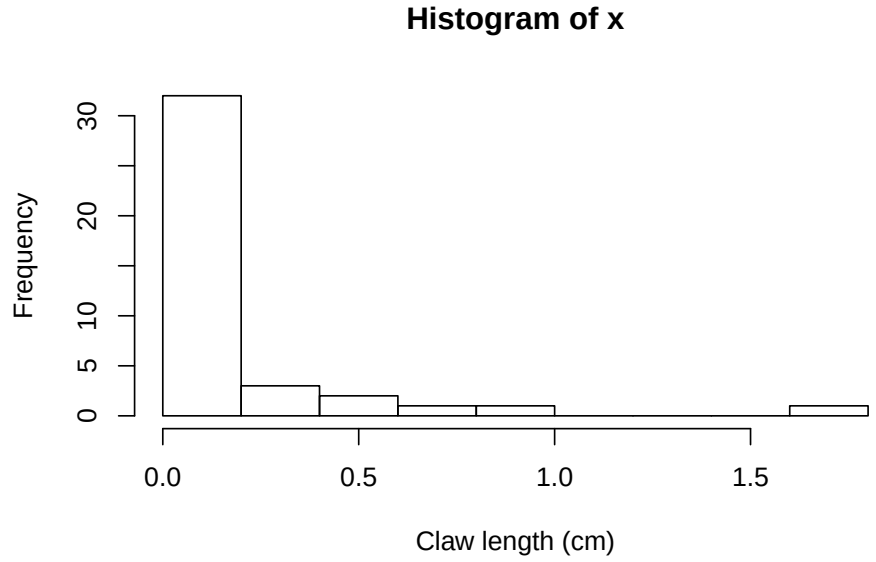
$$f_X(x) = \begin{cases} \frac{2^5}{4!} x^4 e^{-2x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The χ^2_4 density function is given by

$$g_Y(y) = \begin{cases} \frac{1}{2^2}\, y e^{-\frac{y}{2}} & \text{for } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Explain carefully how to use rejection sampling to generate $Gamma(5, 2)$ random variables, using the χ^2_4 density function as the envelope function. Given values $Y = 2.24$ from a χ^2_4 and $U = 0.62$ from a $U[0, 1]$ perform one iteration of the algorithm. ***(9 marks)***

6 (i) The length of the claws of 40 animals are measured and stored in a vector `x`. A histogram is plotted below:



The following analysis is then performed

```
> n <- 10000
> T <- rep(NA, n)
> for(i in 1:n) {
+   S <- sample(x, replace = TRUE)
+   T[i] <- mean(S)
+ }
> var(T)
[1] 0.002450786
```

(a) Explain carefully what procedure has been performed here, and state what the output in the last line represents. *(4 marks)*

(b) Some further analysis is then performed with the output included below

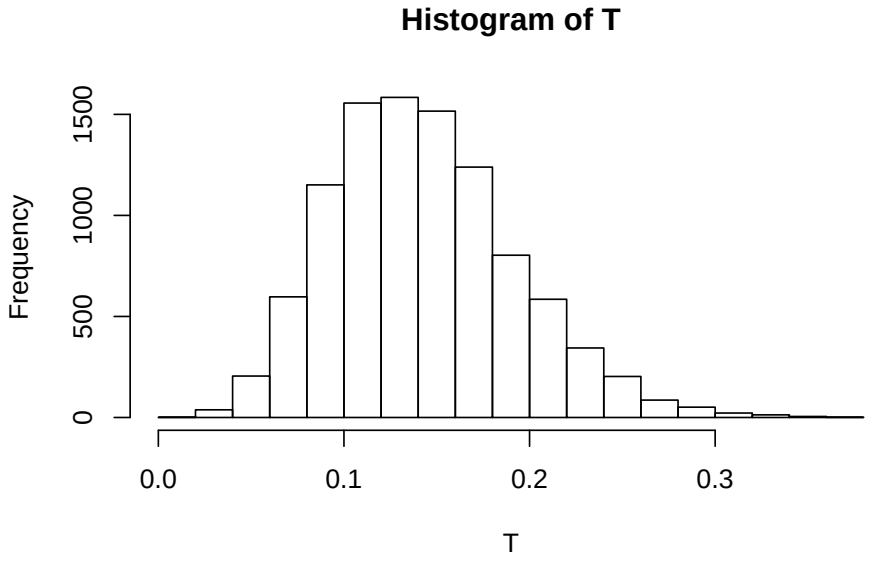
```
> quantile(T, c(0.025, 0.975))
      2.5%      97.5%
0.06034315 0.25056913
> c(mean(T) - 1.96*sd(T), mean(T) + 1.96*sd(T))
[1] 0.04585802 0.23991926
```

State what the outputs represent, and briefly explain the differences between the statistical methods used to obtain the outputs.

(4 marks)

6 (continued)

(c) A histogram of the vector T created by the `R` code is shown below. In light of this plot which of the statistical methods in part (b) would you prefer and why?



(2 marks)

(ii) A random variable Y has an unknown distribution. A sample of 6 observations are taken from the distribution of Y:

$$\{15.9, 19.7, 19.1, 16.7, 23.1, 19.3\}$$

Six random draws from the $U[0, 1]$ distribution are also available:

$$\{0.83, 0.83, 0.96, 0.95, 0.6, 0.26\}$$

(a) Using the six $U[0, 1]$ values, sample one value of a suitable test statistic for use in a randomisation test of the hypothesis $E[Y] = 20$ versus the one-sided alternative that $E[Y] < 20$. ***(7 marks)***

(b) What is the smallest p-value that could be obtained using a permutation test of the hypothesis $E[Y] = \mu$ against a one-sided alternative that $E[Y] < \mu$, given any six observations from the distribution of Y? ***(3 marks)***

End of Question Paper