Data provided: Neave's Tables

MAS472

The University Of Sheffield.

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

MAS472 Computational Inference **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 90.*

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) 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 the mode of the t-distribution, obtain the mean and variance of the importance density. ***(8 marks)***

(b) Given two random draws $Z_1 = -0.52$ and $Z_2 = 0.27$ from a $N(0, 1)$ distribution, obtain two samples from the t_7 distribution via the importance density found above, and calculate the weights of your two sampled values. ***(6 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 $k = 2$.

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

(16 marks)

2 I am performing a sexual health survey and am worried participants may be unwilling to answer certain questions which they consider embarrassing. To try and obtain more truthful responses I use the following strategy. Each interviewee first picks a ball at random from a bag containing 5 red balls and 1 blue ball. If the ball picked is red then the interviewee responds TRUE or FALSE to the statement:

I **have** had a sexually transmitted disease

otherwise, the interviewee responds TRUE or FALSE to the opposite question

I **have never** had a sexually transmitted disease

The interviewer is only given an answer of "TRUE" or "“FALSE"; he is not informed of the colour of the ball selected at any stage. For subject i in the survey, where $i = 1, \ldots, n$, define

$$X_i = \begin{cases} 1 & \text{if the interviewee answers "TRUE",} \\ 0 & \text{if the interviewee answers "FALSE"} \end{cases}$$
$$Y_i = \begin{cases} 1 & \text{if the ball selected is red} \\ 0 & \text{if the ball selected is blue} \end{cases}$$

Define $X = \{X_1, \ldots, X_n\}$, $Y = \{Y_1, \ldots, Y_n\}$ and θ to be the unknown parameter of interest: the proportion of the population who **have** had a sexually transmitted disease.

(i) By splitting the likelihood as

$$P(X, Y|\theta) = \left\{ \prod_{i:Y_i=0} P(Y_i = 0) P(X_i | Y_i = 0, \theta) \right\} \left\{ \prod_{i:Y_i=1} P(Y_i = 1) P(X_i | Y_i = 1, \theta) \right\},$$

show that

$$\begin{aligned} l(\theta; X, Y) =& C + \sum_{i=1}^{n} [(1 - X_i)(1 - Y_i) + X_i Y_i] \log \theta + \\ & \sum_{i=1}^{n} [X_i(1 - Y_i) + Y_i(1 - X_i)] \log(1 - \theta). \end{aligned}$$

where C is a constant that does not depend upon θ ***(7 marks)***

(ii) Show that the maximum likelihood estimate of θ given both X and Y is

$$\hat{\theta} = \frac{\sum_i [X_i Y_i + (1 - X_i)(1 - Y_i)]}{n}.$$

(5 marks)

2 (continued)

(iii) Using Bayes' theorem, show that

$$p_1 = P(Y_i = 1|X_i = 1, \theta) = \frac{5\theta}{4\theta + 1},$$
$$p_0 = P(Y_i = 1|X_i = 0, \theta) = \frac{5(1-\theta)}{5-4\theta}.$$

(7 marks)

(iv) Given X only, suppose the EM algorithm is to be used to obtain the maximum likelihood estimate $\hat{\theta}$ of θ. Let the current estimate of $\hat{\theta}$ be θ_{old}. By maximising

$$Q(\theta|\theta_{old}) = E[l(\theta; X, Y)|X, \theta = \theta_{old}],$$

using your results from parts ii) and iii), show that the new estimate of $\hat{\theta}$ in the EM algorithm is given by

$$\theta_{new} = \frac{sp_1 + (n-s)(1-p_0)}{n},$$

where $p_1 = P(Y_i = 1|X_i = 1, \theta = \theta_{old})$, $p_0 = P(Y_i = 1|X_i = 0, \theta = \theta_{old})$, and s is the total number of observed "TRUE" responses. ***(11 marks)***

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

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

(a) Using integration by parts, show that $E[Y] = \dfrac{1}{\lambda}$ ***(4 marks)***

(b) Explain how to generate a random value of Y given a uniform random number using the inversion method. ***(5 marks)***

(ii) The failure time $t_1, \ldots, t_4$ of four identical light bulbs are claimed to be independent, have mean $1/4$ years and follow an exponential distribution. 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. What is the minimum number of random test statistics required to ensure a size of exactly 0.05? Would you recommend using only this number and why? ***(5 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. ***(5 marks)***

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

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

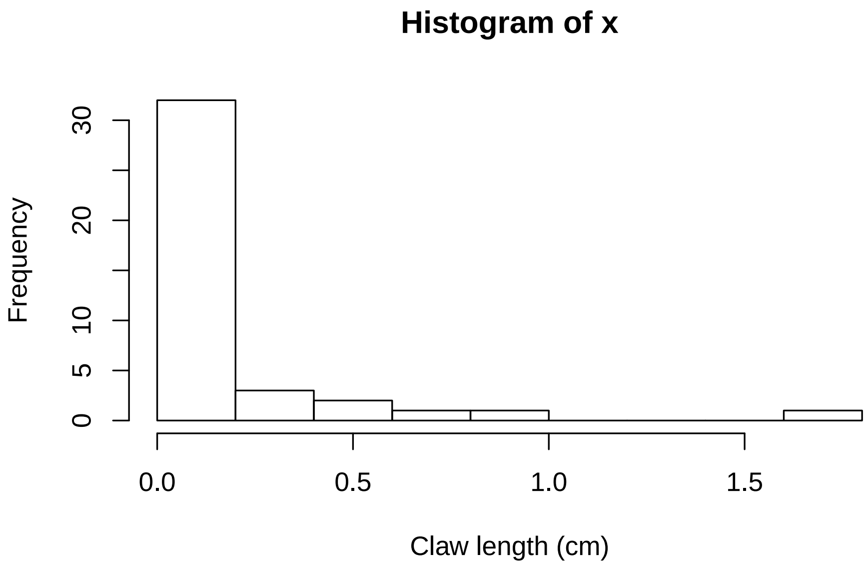
The χ^2_4 density function is given by

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

(a) 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)***

(b) What is the expected number of χ^2_4 random variables needed to generate one gamma random variable. ***(2 marks)***

4 (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) It is suggested that this procedure will converge to the exact answer as `n` is increased. Is this correct? Very briefly, justify your answer. ***(2 marks)***

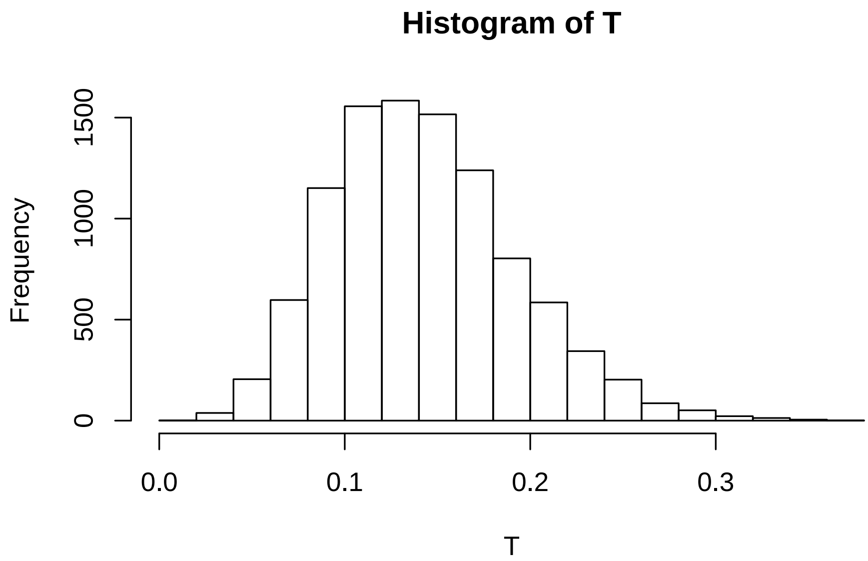
(c) Some further analysis is then performed with the output shown below

```
> 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)***

4 (continued)

(d) 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 (c) would you prefer and why?



(2 marks)

(e) Suppose that it was believed that the observed data `x` arose from a log-normal distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-((\log x - \mu)^2/(2\sigma^2))}$$

where μ and σ are the mean and standard deviation of the logarithm of x respectively.

Why might a log-normal model seem reasonable given the histogram of the observed claw lengths? ***(2 marks)***

(f) Explain carefully how you could adjust the procedure to perform parametric bootstrapping using this log-normal assumption. (It is not necessary to give any R commands) ***(4 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:

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

 Question 4 continued on next page

4 (continued)

(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? When would this happen? ***(5 marks)***

End of Question Paper