Petter Mostad
Mathematical Statistics
Chalmers

MSA100/MVE185 Computer Intensive Statistical Methods

Exam 24 October 2009
Examiner: Petter Mostad, phone 0707163235, visits the exam at 9.30 and at 11.30.
Allowed to use during the exam: Pocket calculator, books, copies, and notes.

1. Eric is working in a company helping people whose car break down on the road. He assumes the number of customers during a time period of t days is Poisson distributed with expectation $t\lambda$, where λ represents the average daily number of customers. In other words, he assumes the probability distribution for y customers in t days is given by

$$\pi(y) = \frac{1}{y!}(t\lambda)^y \exp(-t\lambda).$$

He roughly summarizes his past experience in a discrete prior for λ, where

λ	0.1	0.2	0.3	0.4
Probability	0.1	0.3	0.5	0.1

   (a) If Eric gets 2 customers in 7 days, compute his posterior distribution for λ. (2 points)[1]

   (b) Before making the observations in (a), (i.e., going back to the original prior) compute the probability for Eric to observe zero customers in 7 days. (2 points)

2. You are studying the lengths of females in Sweden in the age group 20-30 years. For a random sample of 10 from this group, you have the following information:

Length	Less than 165 cm	165-175 cm	More than 175 cm
Number of persons	2	5	3

You assume that the lengths in the whole population is normally distributed with mean μ and standard deviation σ. In other words, for a length y, you have the probability density

$$\pi(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y-\mu)^2).$$

Define also the cumulative density function

$$\Phi(z \mid \mu, \sigma) = \int_{-\infty}^{z} \pi(y \mid \mu, \sigma)\, dy.$$

   (a) Assuming that we know that $\sigma = 10$ and using a flat prior for μ, write down a function that is proportional to the posterior distribution for μ, given the information in the table above. (2 points)

   (b) Given such a function, explain one method for simulating from the posterior. (2 points)

[1] The points were not specified in the original exam

(c) Assume that instead we use an improper prior for both μ and σ, so that $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$. Would you guess that the resulting posterior would be proper or improper? Why? What would be your guess if our only data was the number of persons above or below 170 cm? (1 point)

3. A drug company is comparing two drugs used to reduce the chance of heart failure in a group of risk patients. Among 100 persons receiving drug A, 22 experienced heart failure during the test period, while in the group of 100 persons receiving drug B, 12 experienced heart failure. We assume that the number of patients experiencing heart failure is Binomially distributed with probabilities p_A and p_B in the two groups, respectively.

   (a) Assume that the priors for p_A and p_B are both uniform distributions on the interval $[0, 1]$. Prove that the posterior for (p_A, p_B) is a product of two independent Beta distributions, and find their parameters. (If θ has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, its probability density function is $\pi(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$.) (2 points)

   (b) Explain how you can use simulation to compute an approximate 95% credibility interval for the relative risk p_B/p_A of the two drugs. (2 points)

   (c) Explain how you can use simulation to compute the approximate probability that $p_A > p_B$. (1 point)

4. Assume

   `posterior`

   is the name of an R function so that for every value x,

   `posterior(x)`

   is proportional to some posterior density. Assume the proportionality constant is unknown, but that the posterior distribution is known to be *roughly* normal, with expectation 10.2 and standard deviation 5.3. Let

   `func`

   be some other R function taking a single value as input and outputing a single value. Consider the following R code:

```
> v <- rnorm(10000, 10.2, 5.3)
> pv <- rep(0, 10000)
> for (i in 1:10000) pv[i] <- posterior(v[i])
> fv <- rep(0, 10000)
> w <- pv/dnorm(v, 10.2, 5.3)
> for (i in 1:10000) fv[i] <- func(v[i])
> sum(fv*w)/sum(w)
```

   This code computes the approximate value of something: What is it? What is the algorithm implemented in this code called? (2 points)

5. Explain shortly the idea of the Metropolis-Hastings algorithm. (2 points)

6. Assume θ has a prior distribution that is proportional to the function

$$f(\theta) = \exp(-2\theta^2 + \theta - 3).$$

Assume we make two observations, both normally distributed with expectation θ and standard deviation 1, and that these observations are 1 and 4.

   (a) What is the posterior distribution for θ given the data? What is its expectation and variance? (2 points)

   (b) Before we made the observations, what would be the prior predictive distribution for one of the observations[2] assuming the likelihood and prior given above? (2 points)

7. Mary-Ann is examining some data where she has the continuous predictor variable Y, the categorical predictor variable F, and the continuous response variable Y. Her 14 observations can be seen below in the form of a data frame in R. She would like to make a hypothesis test using permutation testing. Her null hypothesis is that, given the value of F, Y is independent of X. She has done the R computations below:[3]

```
> data
           X F         Y
1  4.2770934 A 18.053024
2  3.6325151 C 22.291076
3  2.0884472 C 25.475593
4  5.1032093 B 13.006621
5  1.7058743 A  9.009376
6  4.5813916 A 22.455271
7  5.8215403 B 16.371444
8  4.1420686 C 19.073249
9  5.8678505 B 14.347751
10 1.9898766 B 13.199567
11 0.9651442 B 18.681689
12 2.7350309 A 11.897771
13 2.5694825 C 18.029915
14 2.2625943 A 13.501639
>
> X <- data$X
> Y <- data$Y
> F <- data$F
> mystat <- cor(X, Y, method="kendall")
> simresults <- rep(0, 10000)
> for (i in 1:10000) {

(... see below ...)
```

[2] For clarity, the words “an observation” in the original exam are here replaced by “one of the observations”

[3] In the original exam, the line “F <- data$F” was missing

```
+ simresults[i] <- cor(X, Y, method="kendall")
+ }
> sum(abs(simresults)>abs(mystat))/10000
```

(a) Select one of the four alternatives below for the missing line or lines in the R code above, so that it correctly tests Mary-Anns null hypothesis. Explain your choice. (1 point)

```
Y <- sample(data$Y, replace=TRUE)
```

or

```
+ X <- sample(data$X, replace=FALSE)
```

or

```
+ X[F=="A"] <- sample(data$X[F=="A"], replace=FALSE)
+ X[F=="B"] <- sample(data$X[F=="B"], replace=FALSE)
+ X[F=="C"] <- sample(data$X[F=="C"], replace=FALSE)
```

or

```
+ X[F=="A"] <- sample(data$X[F=="A"], replace=TRUE)
+ X[F=="B"] <- sample(data$X[F=="B"], replace=TRUE)
+ X[F=="C"] <- sample(data$X[F=="C"], replace=TRUE)
```

(b) What is the quantity computed in the last line of the code? (1 point)

8. Assume your likelihood has the form

$$\pi(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y)$$

and that your prior is defined by (for $\alpha > 0, \beta > 0$):

$$\pi(\alpha, \beta) \propto \beta^2 \exp(-\alpha^2 \beta).$$

(a) Show that the prior is in a semi-conjugate familiy to the likelihood. (3 points)

(b) Compute a function that is proportional to the marginal posterior for α. (3 points)