NATIONAL UNIVERSITY OF SINGAPORE
EXAMINATION

ST2334 Probability and Statistics

(Semester 2: AY 2012-2013)

April 2013 Time Allowed: 2 Hours

Instructions for Candidates

1. This examination paper contains **four** (4) questions and comprises **five** (5) printed pages.
2. There are two tables on pages 4 & 5 and results in the tables may be used without proof.
3. Candidates must answer **ALL** questions on the paper.
4. Each question carries 20 marks. The total mark for the paper is 80.
5. Calculators may not be used.
6. Additional statistical tables will not be available.
7. This is a **closed book exam**.

1. (i) Give a definition of a discrete random variable X with appropriate support X. [2 Marks]

   (ii) Let X be a discrete random variable with support X and let $g : \mathsf{X} \to \mathbb{R}$ be a real-valued function. Assuming it exists, write down an expression for the expectation of the random variable $g(X)$. [2 Marks]

   (iii) Let X be a discrete random variable with support X, give a defintion of the moment generating function of X. Call the moment generating function $M(t)$, $t \in \mathsf{T}$; show that

$$\mathbb{E}[X] = \left.\frac{dM(t)}{dt}\right|_{t=0}$$

   assuming that $0 \in \mathsf{T}$ and $\mathbb{E}[X]$ exists. [6 Marks]

   (iv) A manufacturing company produces identical items which are faulty with a probability $p \in (0, 1)$; it is thought that items are faulty independently of each other. The factory will stop manufacturing these items, due to production concerns, the first time that there are $n > 1$ faulty items. Let X be the random variable associated to the first time that there are n faulty items; by using the moment generating function of X, show that **on average** in order to make 1 million items, when $n = 1000$ then we must have $p = 1/1000$. No marks will be awarded if the moment generating function of X is not calculated. [10 Marks]

2. (i) Let X, Y be jointly discrete random variables. Give a definition of the conditional distribution function and the conditional PMF of $Y|X = x$. [3 Marks]

   (ii) Let X, Y be jointly discrete random variables. Give a definition of the conditional expectation of $Y|X = x$. [3 Marks]

   (iii) Let $X, Y \in \mathsf{X} \times \mathsf{Y}$ be jointly discrete random variables and $g : \mathsf{Y} \to \mathbb{R}$. Show that $\mathbb{E}[g(Y)] = \mathbb{E}[\mathbb{E}[g(Y)|X]]$, assuming all expectations exist. [4 Marks]

   (iv) Suppose that $Y \sim \mathcal{P}(\lambda)$ and $X|Y = y \sim \mathcal{B}(y, p)$.

   (a) Find the moment generating function of $X|Y$. [3 Marks]

   (b) Find the moment generating function of X and hence $\mathbb{E}[X]$. [7 Marks]

3. (i) Consider continuous random variables X, Y. What does it mean for them to be jointly continuous, with joint PDF $f(x, y)$? [3 Marks]

 (ii) Let X, Y be jointly continuous random variables, with joint PDF $f(x, y)$. Give a definition of the marginal PDFs of X and Y. [3 Marks]

 (iii) Let X, Y be jointly continuous random variables, with joint PDF $f(x, y)$. Under what conditions are the random variables independent? [4 Marks].

 (iv) Let $\mathsf{Z} = (\mathbb{R}^+)^2 = \mathsf{X} \times \mathsf{Y}$ and

$$f(x, y) = \lambda^2 e^{-\lambda(x+y)} \quad (x, y) \in \mathsf{Z}, \lambda > 0.$$

 (a) Find the marginal PDFs of X and Y; are X and Y independent? [5 Marks]

 (b) Calculate the probability that $X > Y$. [5 Marks]

4. (i) Let $X_1, \ldots, X_n$ be mutually independent, with $X_i \sim F_\theta$ and associated PDF/PMF f_θ, where θ is an unknown and possibly multi-dimensional parameter. Give the approach detailed in lectures to obtain the maximum likelihood estimator (MLE) $\widehat{\theta}_n$ of θ. [5 Marks]

 (ii) Let $X_1, \ldots, X_n$ be mutually independent, with $X_i \sim F(\cdot|\theta)$ and associated PDF/PMF $f(\cdot|\theta)$, where θ is an unknown and possibly multi-dimensional parameter and we assume that it is a random variable. Give Bayes theorem and discuss how one could estimate the parameter. Some minor contrast with the MLE should also be given. [5 Marks]

 (iii) Suppose one observes n independent and identically distributed exponential random variables, $\mathcal{E}(\lambda)$, $X_1, \ldots, X_n$, with $\lambda > 0$ unknown.

 (a) Compute the maximum likelihood estimator. [5 Marks]

 (b) Now, suppose one takes Bayesian perspective an assumes *a-priori* that $\lambda \sim \mathcal{G}(a, b)$ for some $a, b > 0$ known. Find the posterior mean, that is

$$\int_0^\infty \lambda \pi(\lambda | x_1, \ldots, x_n) d\lambda$$

 where $\pi(\lambda | x_1, \ldots, x_n)$ is the posterior PDF of λ. Contrast this estimate with the MLE in (a). [5 Marks]

END OF PAPER

	Support X	Par.	PMF	CDF	$\mathbb{E}[X]$	$\mathbb{V}\mathrm{ar}[X]$	MGF
$\mathcal{B}(1,p)$	$\{0,1\}$	$p \in (0,1)$	$p^x(1-p)^{1-x}$		p	$p(1-p)$	$(1-p)+pe^t$
$\mathcal{B}(n,p)$	$\{0,1,\ldots,n\}$	$p \in (0,1)$, $n \in \mathbb{Z}^+$	$\binom{n}{x}p^x(1-p)^{1-x}$		np	$np(1-p)$	$((1-p)+pe^t)^n$
$\mathcal{P}(\lambda)$	$\{0,1,2,\ldots\}$	$\lambda \in \mathbb{R}^+$	$\frac{\lambda^x e^{-\lambda}}{x!}$		λ	λ	$\exp\{\lambda(e^t-1)\}$
$\mathcal{G}e(p)$	$\{1,2,\ldots\}$	$p \in (0,1)$	$(1-p)^{x-1}p$	$1-q^x$	$1/p$	$(1-p)/p^2$	$\frac{pe^t}{1-e^t(1-p)}$
$\mathcal{N}e(n,p)$	$\{n,n+1,\ldots\}$	$p \in (0,1), n \in \mathbb{Z}^+$	$\binom{x-1}{n-1}(1-p)^{x-n}p^n$		n/p	$n(1-p)/p^2$	$\left(\frac{pe^t}{1-e^t(1-p)}\right)^n$

Table 1: Table of Discrete Distributions. Note that $q = 1 - p$.

	Support X	Par.	PDF	CDF	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$	MGF
$\mathcal{U}_{[a,b]}$	$[a,b]$	$-\infty < a < b < \infty$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$(a+b)/2$	$(b-a)^2/12$	$\frac{e^{bt}-e^{at}}{t(b-a)}$
$\mathcal{E}(\lambda)$	$\mathbb{R}^+$	$\lambda \in \mathbb{R}^+$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda - t}$
$\mathcal{G}(a,b)$	$\mathbb{R}^+$	$a, b \in \mathbb{R}^+$	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$		a/b	a/b^2	$(\frac{b}{b-t})^a$
$\mathcal{N}(a,b)$	$\mathbb{R}$	$(a,b) \in \mathbb{R} \times \mathbb{R}^+$	$\frac{1}{\sqrt{2\pi b}} e^{-\frac{1}{2b}(x-a)^2}$		a	b	$e^{at+bt^2/2}$
$\mathcal{B}e(a,b)$	$[0,1]$	$(a,b) \in (\mathbb{R}^+)^2$	$B(a,b)^{-1} x^{a-1} (1-x)^{b-1}$		$a/(a+b)$	$\frac{ab}{(a+b)^2(a+b+1)}$	

Table 2: Table of Continuous Distributions. Note that $B(a,b)^{-1} = \Gamma(a+b)/[\Gamma(a)\Gamma(b)]$.

NUS EXAMINATIONS (STATISTICS)

April 2013

ST2334

Probability and Statistics (Solutions)

Setter's signature	Checker's signature	Editor's signature
....	....	....

1. (i) A random variable X is said to be discrete if it takes values in some countable subset $\mathsf{X} = \{x_1, x_2, \dots\}$ of $\mathbb{R}$. $\boxed{2}$

(ii) Suppose that X has PMF $f(x)$ on X, then

$$\mathbb{E}[g(X)] = \sum_{x \in \mathsf{X}} g(x) f(x).$$

$\boxed{2}$

(iii) For a discrete random variable X the **moment generating function** (MGF) is

$$M(t) = \mathbb{E}[e^{Xt}] = \sum_{x \in \mathsf{X}} e^{xt} f(x) \quad t \in \mathsf{T}$$

where T is the set of t for which $\sum_{\mathsf{X}} e^{xt} f(x) < \infty$. Now, we have for $t \in \mathsf{T}$

$$\frac{dM(t)}{dt} = \frac{d}{dt} \sum_{x \in \mathsf{X}} e^{xt} f(x) = \sum_{x \in \mathsf{X}} x e^{xt} f(x).$$

Assuming $0 \in \mathsf{T}$, we have

$$\left.\frac{dM(t)}{dt}\right|_{t=0} = \sum_{x \in \mathsf{X}} x f(x) = \mathbb{E}[X].$$

$\boxed{6}$

(iv) X is clearly a negative binomial random variable with parameters n, p. We first compute the MGF:

$$\begin{aligned} M(t) &= \sum_{x=n}^{\infty} e^{xt} \binom{x-1}{n-1} p^n (1-p)^{x-n} \\ &= \Big(\frac{p}{1-p}\Big)^n \sum_{x=n}^{\infty} \binom{x-1}{n-1} (e^t(1-p))^x. \end{aligned}$$

Let

$$t \in \mathsf{T} = \{t \in \mathbb{R} : t < -\log(1-p)\}.$$

In this scenario

$$0 < e^t(1-p) < 1.$$

Then setting $p' = e^t(1-p)$ we have

$$\begin{aligned} M(t) &= \Big(\frac{p}{1-p}\Big)^n \Big(\frac{p'}{(1-p')}\Big)^n \sum_{x=n}^{\infty} \binom{x-1}{n-1} (p')^{x-n} (1-p')^n \\ &= \Big(\frac{pp'}{(1-p)(1-p')}\Big)^n. \end{aligned}$$

The last line follows as the summation is 1; it is the sum of a negative binomial PMF with parameters $n, 1-p'$.

Then on using the definition of p', for $t \in \mathsf{T}$

$$M(t) = \Big(\frac{pe^t}{1-(1-p)e^t}\Big)^n.$$

Now, when $n = 1000$ we want

$$\mathbb{E}[X] = 1000000.$$

As we are to use the moment generating function, we differentiate w.r.t. t and then set $t = 0$ to find the expectation. We have

$$\frac{dM(t)}{dt} = nM(t)\Big(1 + \frac{(1-p)e^t}{1-(1-p)e^t}\Big)$$

Setting $t = 0$ we have

$$\mathbb{E}[X] = n/p.$$

Thus we want:

$$p = \frac{1000}{1000000} = \frac{1}{1000}.$$

$\boxed{10}$

2. (i) The conditional distribution function of Y given X, written $F_{Y|x}(\cdot|x)$, is defined by

$$F_{y|x}(y|x) = \mathbb{P}(Y \leq y|X = x)$$

for any x with $\mathbb{P}(X = x) > 0$. The conditional PMF of Y given $X = x$ is defined by

$$f(y|x) = \mathbb{P}(Y = y|X = x)$$

when x is such that $\mathbb{P}(X = x) > 0$. $\boxed{3}$

(ii) The **conditional expectation** of a random variable Y, given $X = x$ is

$$\mathbb{E}[Y|X = x] = \sum_y y f(y|x)$$

given that the conditional PMF is well-defined. $\boxed{3}$

(iii) We have

$$\begin{aligned}
\mathbb{E}[g(Y)] &= \sum_y g(y)f(y) \\
&= \sum_{(x,y)\in \mathsf{Z}} g(y)f(x,y) \\
&= \sum_{(x,y)\in \mathsf{Z}} g(y)f(y|x)f(x) \\
&= \sum_x [\sum_y g(y)f(y|x)]f(x) \\
&= \mathbb{E}[\mathbb{E}[g(Y)|X]].
\end{aligned}$$

$\boxed{4}$

(iv) (a) $X|Y \sim \mathcal{B}(y,p)$; thus

$$M_{X|Y}(t) = (q + pe^t)^y$$

where $q = 1 - p$ (from Tables). $\boxed{3}$

(b) We have

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{Xt}] \\
&= \mathbb{E}[\mathbb{E}[e^{Xt}|Y]] \\
&= \mathbb{E}[(q + pe^t)^Y] \\
&= \sum_{y=0}^{\infty} (q + pe^t)^y \frac{\lambda^y e^{-\lambda}}{y!} \\
&= \exp\{\lambda((q + pe^t) - 1)\}.
\end{aligned}$$

Differentiating

$$\frac{dM_X(t)}{dt} = \exp\{\lambda((q + pe^t) - 1)\}\lambda p.$$

Thus on setting $t = 0$; $\mathbb{E}[X] = \lambda p$. $\boxed{7}$

3. (i) The random variables are jointly continuous with joint PDF $f : \mathbb{R}^2 \to [0, \infty)$ if

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v) du dv$$

for each $(x,y) \in \mathbb{R}^2$. $\boxed{3}$

(ii) The **marginal density functions** of X and Y

$$f(x) = \int_{-\infty}^{\infty} f(x,y) dy \quad f(y) = \int_{-\infty}^{\infty} f(x,y) dx.$$

$\boxed{3}$

(iii) The random variables X and Y are independent if and only if

$$F(x,y) = F(x)F(y)$$

or equivalently

$$f(x,y) = f(x)f(y).$$

$\boxed{4}$

(iv) (a) One has

$$f(x) = \lambda e^{-\lambda x} \; x \in \mathsf{X} \quad f(y) = \lambda e^{-\lambda y} \; y \in \mathsf{Y}$$

so clearly X and Y are independent. $\boxed{5}$

(b)

$$\begin{aligned}
\mathbb{P}(X > Y) &= \int_0^{\infty} \int_0^{x} f(x,y) dy dx \\
&= \int_0^{\infty} \int_0^{x} \lambda e^{-\lambda y} dy \lambda e^{-\lambda x} dx \\
&= \int_0^{\infty} \Big[- e^{-\lambda y} \Big]_0^{x} \lambda e^{-\lambda x} dx \\
&= \int_0^{\infty} (1 - e^{-\lambda x}) \lambda e^{-\lambda x} dx \\
&= \Big[- e^{-\lambda x} + \frac{1}{2} e^{-2\lambda x} \Big]_0^{\infty} \\
&= 1 - \frac{1}{2} = \frac{1}{2}.
\end{aligned}$$

$\boxed{5}$

4. (i) The joint pmf/pdf is:

$$f_\theta(x_1,\ldots,x_n) = f_\theta(x_1)\times f_\theta(x_2)\times\cdots\times f_\theta(x_n) = \prod_{i=1}^n f_\theta(x_i).$$

We call $f_\theta(x_1,\ldots,x_n)$ the *likelihood* of the data. As maximizing a function is equivalent to maximizing a monotonic increasing transformation of the function, we often work with the *log-likelihood*:

$$l_\theta(x_1,\ldots,x_n) = \log\Big(f_\theta(x_1,\ldots,x_n)\Big) = \sum_{i=1}^n \log\Big(f_\theta(x_i)\Big).$$

If Θ is some continuous space (as it generally is for our examples) and $\theta = (\theta_1,\ldots,\theta_d)$, then we can compute the gradient vector:

$$\nabla l_\theta(x_1,\ldots,x_n) = \Big(\frac{\partial l_\theta(x_1,\ldots,x_n)}{\partial\theta_1},\ldots,\frac{\partial l_\theta(x_1,\ldots,x_n)}{\partial\theta_d}\Big)$$

and we would like to solve, for $\theta\in\Theta$ (below 0 is the $d-$dimensional vector of zeros)

$$\nabla l_\theta(x_1,\ldots,x_n) = 0. \tag{1}$$

The solution of this equation (assuming it exists) is a maximum if the *hessian matrix* is negative definite:

$$H(\theta) := \begin{bmatrix} \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_1^2} & \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_1\partial\theta_d} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_d\partial\theta_1} & \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_d\partial\theta_2} & \cdots & \frac{\partial^2 l_\theta(x_1,\ldots,x_n)}{\partial\theta_d^2} \end{bmatrix}.$$

If the d numbers $\lambda_1,\ldots,\lambda_d$ which solve$|\lambda I_d - H(\theta)| = 0$, with I_d the $d\times d$ identity matrix, are all negative, then θ is a local maximum of $l_\theta(x_1,\ldots,x_n)$. If $d=1$ then this just boils down to checking whether the second derivative of the log-likelihood is negative at the solution of (1).

Thus in summary, the approach we employ is as follows:

1. Compute the likelihood $f_\theta(x_1,\ldots,x_n)$.
2. Compute the log-likelihood $l_\theta(x_1,\ldots,x_n)$ and its gradient vector $\nabla l_\theta(x_1,\ldots,x_n)$.
3. Solve $\nabla l_\theta(x_1,\ldots,x_n) = 0$, with respect to $\theta\in\Theta$, call this solution $\widetilde{\theta}_n$ (we are assuming there is only one $\widetilde{\theta}_n$).
4. If $H(\widetilde{\theta}_n)$ is negative definite, then $\widehat{\theta}_n = \widetilde{\theta}_n$.

In general, point 3. may not be possible analytically (so for example, one can use Newton's method). H

$\boxed{5}$

(ii) Thus we have that the joint pmf/pdf of $X_1, \dots, X_n|\theta$ is:

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

The main key behind Bayesian statistics is the choice of a *prior* probability distribution for the parameter $\theta \in \Theta$. That is, Bayesian statisticians specify a probability distribution on the parameter θ *before* the data are observed. This probability distribution is supposed to reflect the information one might have before seeing the observations. Throughout, we will write the prior pmf/pdf as $\pi(\theta)$. Now the way in which Bayesian inference works is to update the prior beliefs on θ via the posterior pmf/pdf. That is, 'in the light of the data' the distributional properties of the prior are updated. This is achieved by Bayes theorem; the *posterior* pmf/pdf is:

$$\pi(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta)\pi(\theta)}{f(x_1, \dots, x_n)}$$

where

$$f(x_1, \dots, x_n) = \int_{\Theta} f(x_1, \dots, x_n|\theta)\pi(\theta)d\theta$$

if θ is continuous and, if θ is discrete:

$$f(x_1, \dots, x_n) = \sum_{\theta \in \Theta} f(x_1, \dots, x_n|\theta)\pi(\theta).$$

For a Bayesian statistician, the posterior is the 'final answer', in that all statistical inference should be associated to the posterior. For example, if one is interested in estimating θ then one can use the posterior mean:

$$\mathbb{E}[\Theta|x_1, \dots, x_n] = \int_{\Theta} \theta\pi(\theta|x_1, \dots, x_n)d\theta.$$

The posterior distribution is much 'richer' than the MLE, in the sense that one now has a whole distribution which reflects the parameter, instead of a point estimate. What also might be apparent, is the fact that the posterior is perhaps difficult to calculate. 5

(iii) (a) Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{E}(\lambda)$ random variables. Let us compute the MLE of $\lambda = \theta$, given observations $x_1, \dots, x_n$. First, we have

$$\begin{aligned} f_\lambda(x_1, \dots, x_n) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\}. \end{aligned}$$

Second the log-likelihood is:

$$l_\lambda(x_1, \dots, x_n) = \log(f_\lambda(x_1, \dots, x_n)) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i.$$

The gradient vector is a derivative:

$$\frac{dl_\lambda(x_1,\dots,x_n)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Thirdly

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

so

$$\widetilde{\lambda}_n = (\frac{1}{n}\sum_{i=1}^{n} x_i)^{-1}.$$

Fourthly,

$$\frac{d^2 l_\lambda(x_1,\dots,x_n)}{d\lambda^2} = -\frac{n}{\lambda^2} < 0.$$

Thus

$$\widehat{\lambda}_n = (\frac{1}{n}\sum_{i=1}^{n} x_i)^{-1}.$$

$\boxed{5}$

(b) Now for the Bayesian model: Here we have that

$$\begin{aligned}
f(x_1,\dots,x_n) &= \int_0^\infty \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\} \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}d\lambda \\
&= \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{n+a-1}\exp\{-\lambda[\sum_{i=1}^{n} x_i + b]\}d\lambda \\
&= \frac{b^a}{\Gamma(a)}\Big(\frac{1}{[\sum_{i=1}^{n} x_i + b]}\Big)^{n+a}\int_0^\infty u^{n+a-1}e^{-u}du \\
&= \frac{b^a}{\Gamma(a)}\Big(\frac{1}{[\sum_{i=1}^{n} x_i + b]}\Big)^{n+a}\Gamma(n+a).
\end{aligned}$$

So, as:

$$f(x_1,\dots,x_n|\lambda)\pi(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{n+a-1}\exp\{-\lambda[\sum_{i=1}^{n} x_i + b]\}$$

we have:

$$\pi(\lambda|x_1,\dots,x_n) = \frac{\lambda^{n+a-1}\exp\{-\lambda[\sum_{i=1}^{n} x_i + b]\}}{\Big(\frac{1}{[\sum_{i=1}^{n} x_i + b]}\Big)^{n+a}\Gamma(n+a)}$$

i.e.

$$\lambda|x_1,\dots,x_n \sim \mathcal{G}(n+a, b+\sum_{i=1}^{n} x_i).$$

So, for example:

$$\mathbb{E}[\Lambda|x_1,\dots,x_n] = \frac{n+a}{b+\sum_{i=1}^{n} x_i}.$$

In comparison to the MLE, we see that the posterior mean and MLE correspond as $a, b \to 0$. $\boxed{5}$