Student ID

Semester 2 Assessment, 2014

Department of Mathematics and Statistics

MAST20005 Statistics

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

Common content with: MAST90058

This paper consists of 7 pages (including this page)

Authorised materials:

- The following materials are authorised: Calculator of any kind, one double-sided A4 sheet of handwritten notes.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.

  All answers are to be written in your answer booklet. All questions may be attempted. The total number of marks available is 70.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.
- Students may bring one double-sided A4 sheet of handwritten notes into the examination room.
- Students may take this paper with them at the end of the exam.
- Textbooks are not allowed.

Blank page (ignored in page numbering)

Question 1 (12 marks) Let $X_1, \dots, X_n$ be a random sample from the probability density function (pdf):

$$f(x;\theta) = \frac{x}{\theta^2} e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$$

(a) Determine a sufficient statistic for θ.

(b) Write down the log-likelihood function, $\ell(\theta)$, and the score function, $s(\theta)$.

(c) Determine the maximum likelihood estimator of θ.

(d) Give the Crámer-Rao lower bound of unbiased estimators of θ. Hint: If X follows a Gamma(α, β) distribution with pdf $(\beta^\alpha x^{\alpha-1} e^{-x/\beta})/\Gamma(\alpha)$ $(x, \alpha, \beta > 0)$, then $E[X] = \alpha\beta$.

(e) A random sample of size $n = 35$ on X where X has the density $f(x;\theta)$ gave $\overline{x} = 10.5$. Determine the maximum likelihood estimate of θ and an approximate 95% confidence interval for θ. Some R output that may help.

```
>  z <- c(0.95,0.975,0.99,0.995)
> qnorm(z)
[1] 1.644854 1.959964 2.326348 2.575829
```

Question 2 (4 marks) Let $X_1, \dots, X_n$ be a random sample from the Gamma(α, β) distribution with probability density function (pdf) $(\beta^\alpha x^{\alpha-1} e^{-x/\beta})/\Gamma(\alpha)$ $(x, \alpha, \beta > 0)$. Recall that if $X \sim$ Gamma(α, β), then $E[X] = \alpha\beta$ and $Var(X) = \alpha\beta^2$.

(a) Derive a estimators for α and β using the method of moments.

(b) Find point estimates of α and β from the following observations on X:

```
3.51 3.27 4.90 5.27 4.25
```

Question 3 (6 marks) The number of patients arriving at a minor injuries clinic in 10 half-hour intervals are recorded. It is supposed that, given the value of a parameter λ, the number X_j arriving in interval j has a Poisson distribution $X_j \sim$ Pois(λ) with pmf

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \{0, 1, \dots\}, \ \lambda > 0,$$

and X_j is independent of X_j, for $j \neq k$. The prior distribution for λ is taken to be Gamma($\alpha = 2, \beta = 2$) distribution (recall that if X follows a Gamma(α, β) distribution with pdf $(\beta^\alpha x^{\alpha-1} e^{-x/\beta})/\Gamma(\alpha)$ $(x, \alpha, \beta > 0)$, then $E[X] = \alpha\beta$.). Find the posterior distribution of λ using the following observations on X:

```
9 12 16 12 16 11 18 13 12 19
```

and give the bayesian point estimate for λ under the squared loss function $[w(y) - \lambda]^2$.

Question 4 (12 marks) Let $X_1, \ldots, X_n$ be a random sample from the Exponential distribution $\text{Exp}(\theta)$ with pdf:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0, \quad \theta > 0,$$

and cumulative distribution function $F(x;\theta) = 1 - e^{-x/\theta}$.

(a) Derive the cumulative distribution function of $W = \min_{i=1,\ldots,n} X_i$ and then name the distribution of W.

(b) Use W to define an unbiased estimator for θ and show that such an estimator does not attain the Crámer-Rao lower bound for $n > 1$. *Hint: recall that the maximum likelihood estimator for θ is $\sum_{i=1}^{n} X_i/n$.*

(c) Use the statistic W to construct a $100(1-\alpha)\%$ confidence interval for θ.

(d) Let X be the lifetime (in months) of electric bulbs made by company XYZ. Consider the following sample of $n = 5$ observations on X:

```
0.16  9.34  54.93  8.75  27.47
```

Use the result in (c) to find a 95% confidence interval for the true mean lifetime

Question 5 (10 marks) An eating attitude test (EAT) was administered to both a sample of female models and a control group of females, resulting in the following summary statistics.

	Sample size	Sample mean	Sample standard deviation
Models	30	8.63	4.1
Controls	30	10.97	5.9

You may assume both samples came from normal populations.

(a) Is there sufficient evidence to justify claiming that a difference exists in the mean EAT score between models and controls? Assume that the two populations have equal variances and use a test with significance level $\alpha = 0.05$ and clearly state your null and alternative hypotheses, test statistic and its distribution.

(b) Give an approximate p-value for the test in (a) (provide an interval in which the actual p-value for the test in (a) lies).

(c) Is there evidence that the population variances differ between models and controls? Justify your answer by giving an appropriate confidence interval.

The following R output may be useful.

```
> s=c(0.005,0.01,0.025,0.05, 0.950, 0.975, 0.990, 0.995)
> qt(s, 58)
 -2.66 -2.39 -2.00 -1.67  1.67  2.00  2.39  2.66
> qnorm(s)
 -2.58 -2.33 -1.96 -1.64  1.64  1.96  2.33  2.58
> qt(s, 29)
 -2.76 -2.46 -2.05 -1.70  1.70  2.05  2.46  2.76
> qf(s, 29, 29)
```

```
  0.37  0.41  0.48  0.54  1.86  2.10  2.42  2.67
> qf(s, 58, 58)
  0.50  0.54  0.59  0.65  1.55  1.68  1.86  1.99
```

Question 6 (6 marks) A researcher has developed a theoretical model for the number of DNA mutations in humans, X, occurring in the DNA region K. She predicted that X follows a Poisson distribution with mean $E(X) = 2$ (the corresponding pmf is $2^x e^{-2}/x!$, $x = 0, 1, \dots$). To test her claim, she collected observations based on 100 healthy subjects and obtained the following .

Number of mutations:	0	1	2	3	4	≥ 5
Observed frequency:	10	30	28	15	9	8

(a) Is there evidence in these data suggesting that the model developed by the researcher is faulty? Answer by carrying out an appropriate hypothesis test at the 0.05 significance level. In your test, state clearly hypotheses, the observed value of test statistic and its distribution, and a conclusion in the context of these data.

(b) Explain how the distribution of the statistic in part (a) changes when the null hypothesis states that the data are generated by a generic Poisson distribution (i.e. no information is given about the mean λ).

Some R output that may be useful is

```
> dpois(0:8, 2)
 0.14 0.27 0.27 0.18 0.09 0.04 0.01 0.00 0.00
> ppois(0:8, 2)
 0.14 0.41 0.68 0.86 0.95 0.98 1.00 1.00 1.00
 > qchisq(c(0.90, 0.95, 0.975), 1)
[1] 2.705543 3.841459 5.023886
> qchisq(c(0.90, 0.95, 0.975), 2)
[1] 4.605170 5.991465 7.377759
> qchisq(c(0.90, 0.95, 0.975), 5)
[1]  9.236357 11.070498 12.832502
> qchisq(c(0.90, 0.95, 0.975), 4)
[1]  7.779440  9.487729 11.143287
```

Question 7 (8 marks) Researchers were interested in whether chocolate milk might be an equivalent (or even better) recovery drink for athletes. So they took 16 cyclists and had them do a strenuous interval workout. Then they were allowed a four-hour recovery period where eight were randomly assigned to drink chocolate milk and the others were randomly assigned to drink a carb-replacement drink. After the recovery period, they were asked to ride until exhaustion, and their time (in minutes) to exhaustion was measured.

Chocolate milk (X):	49.1	51.0	50.9	53.6	50.6	49.8	50.0	19.7
Carb-replacement (Y):	59.0	20.3	19.8	22.2	20.8	19.6	20.0	22.9

(a) Use the sign test to determine if there is evidence that the median exhaustion time for chocolate milk drinkers is larger than 50 minutes. Use $\alpha = 0.05$ and clearly state your hypotheses.

Some R output that may be useful is

```
> pbinom(0:8, 8 ,.5)
  0.00 0.04 0.14 0.36 0.64 0.86 0.96 1.00 1.00
> dbinom(0:8, 8 ,.5)
  0.00 0.03 0.11 0.22 0.27 0.22 0.11 0.03 0.00
```

(b) Take $\alpha = 0.05$ and use the statistic W – defined by the sum of the ranks of the observations of Y (carb-replacement drinkers) in the combined sample – to test $H_0 : m_x = m_y$ against $H_1 : m_x \neq m_y$, where m_x and m_y denote the true medians of X and Y, respectively. Recall that when $Z \sim N(0, 1)$, we have $P(Z \leq 1.645) \approx 0.95$, $P(Z \leq 1.96) \approx 0.975$

Question 8 (12 marks) The following data table gives observations on total acidity of coal samples of three different types, with determinations made using three different concentrations of ethanolic NaOH (measured in moles per litres, N = mol/L).

		Type of coal		
		1	2	3
	.404N	8.27, 8.17	8.66, 8.61	8.14, 7.96
NaOH concentration	.626N	8.03, 8.21	8.42, 8.58	8.02, 7.89
	.786N	8.60, 8.20	8.61, 8.76	8.13, 8.07

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, i = 1, 2, 3, \ j = 1, 2, 3, \ k = 1, 2.$$

Use the attached R output to answer the following questions. You should adopt the convention that the interaction sum of squares is incorporated into the error sum of squares if the interaction is not significant.

(a) Consider testing $H_0 : \gamma_{ij} = 0$, $i = 1, 2, 3$, $j = 1, 2, 3$ (no interaction), at the 5% level of significance. What is the value of the test statistic, its distribution when H_0 is true and the p-value of the test?

(b) Consider testing 1) $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ (no row effect) and 2) $H_0 : \beta_1 = \beta_2 = \beta_3$ (no column effect) at the 5% level of significance. What are the values of the test statistics, their distribution when H_0 is true and the p-value of the tests?

(c) Give a point estimate of the variance σ^2.

(d) Briefly state your conclusions from (a) and (b) in the context of the data at hand.

(e) Let $\overline{X}_{1\cdot} = 8.30$, $\overline{X}_{2\cdot} = 8.19$ and $\overline{X}_{3\cdot} = 8.40$ be the row means for the NaOH tratment groups. To compare row treatment 1 (.404N NaOH concentration) to the other two treatments (.626N and .786N), derive a test statistic for null hypothesis $H_0 : \alpha_1 = (\alpha_2/2 + \alpha_3/2)$ against the alternative $H_1 : \alpha_1 \neq (\alpha_2/2 + \alpha_3/2)$. Give the test statistic and a rejection region for H_0. In your test, consider unknown but equal group variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

```
> x = c(8.27, 8.17 , 8.66, 8.61 , 8.14, 7.96, 8.03, 8.21 ,
+ 8.42, 8.58 , 8.02, 7.89,8.60, 8.20 , 8.61, 8.76 , 8.13, 8.07)
> A <- as.factor(rep(1:3,6))
> B <- as.factor(rep(1:3,each=6))
> tapply(x, list(A,B), mean)
      1     2     3
1 8.220 8.120 8.400
2 8.635 8.500 8.685
3 8.050 7.955 8.100
> A  <- as.factor(A);
> B  <- as.factor(B);
> M1 <- lm(x˜A + B + A*B);
> M2 <- lm(x˜A+B);
> anova(M1)
Analysis of Variance Table

Response: x
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          2 1.00241 0.50121 29.4923 0.0001117 ***
B          2 0.12431 0.06216  3.6574 0.0687816 .
A:B        4 0.01456 0.00364  0.2141 0.9240134
Residuals  9 0.15295 0.01699
---
Signif. codes:  0 ***  0.001 ** 0.01 * 0.05 .

> anova(M2)
Analysis of Variance Table

Response: x
          Df  Sum Sq Mean Sq F value   Pr(>F)
A          2 1.00241 0.50121 38.8982 3.26e-06 ***
B          2 0.12431 0.06216  4.8239   0.0271 *
Residuals 13 0.16751 0.01289
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

> qt(c(0.975), 18)
 2.100922
> qt(c(0.975), 15)
 2.13145
> qt(c(0.975), 17)
 2.109816
```

End of Exam—Total Available Marks = 70