Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2017	MATH4071-WE01

Title: Topics in Statistics IV

Time Allowed:	3 hours	
Additional Material provided:	Tables: Normal, t-distribution, F-distribution, χ^2-distribution; Graph paper.	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio fx-83 GTPLUS or Casio fx-85 GTPLUS.
Visiting Students may use dictionaries: No		

Instructions to Candidates:	Credit will be given for: the best **TWO** answers from Section A, the best **THREE** answers from Section B, **AND** the answer to the question in Section C. Questions in Section B and C carry **TWICE** as many marks as those in Section A.

Revision:	

SECTION A

1. For a continuous probability density function $p(y;\theta)$ with parameter vector θ and using the notation ∂_θ for partial differentiation with respect to the parameters, show that

$$\mathrm{E}[\,\partial_\theta \log p(Y;\theta)\,] = 0$$

and

$$\mathrm{Var}[\,\partial_\theta \log p(Y;\theta)\,] = -\mathrm{E}[\,\partial_\theta^2 \log p(Y;\theta)\,]$$

when the derivatives are evaluated at the same value of θ as used to compute the expectations.

Now consider maximum likelihood estimation of θ based on a random sample. Derive the asymptotic distribution of the maximum likelihood estimator, $\hat{\theta}$. You may assume that, asymptotically, $\hat{\theta} \to \theta_0$ and may treat as exact the second-order Taylor approximation to the log-likelihood:

$$L(\theta) \approx L(\theta_0) + D^{\mathrm{T}}(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^{\mathrm{T}} H (\theta - \theta_0)$$

where θ_0 is the true parameter value, D is the gradient $L'(\theta_0)$ and H the Hessian $L''(\theta_0)$.

You may use without proof any necessary standard properties of the multivariate normal distribution. You may also use, but should refer to when used, any standard theorems from probability such as the law of large numbers and the central limit theorem.

2. In a two-parameter statistical model, numerical maximisation of the log-likelihood function $L(\theta)$ gave

$$\hat{\theta} = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \qquad \text{and} \qquad L''(\hat{\theta}) = \begin{pmatrix} -9/5 & -2/5 \\ -2/5 & -1/5 \end{pmatrix}$$

Explaining your working carefully, make a detailed sketch of the 95% Wald confidence region for θ. Your sketch should include axes in the units used to provide the numbers in the question.

3. The full (saturated) log-linear model for a three-way contingency table having R rows, C columns and S slices is

$$\log p_{ijk} = \eta_{ijk} = \beta + \beta_i^{(1)} + \beta_j^{(2)} + \beta_k^{(3)} + \beta_{ij}^{(12)} + \beta_{ik}^{(13)} + \beta_{jk}^{(23)} + \beta_{ijk}^{(123)}$$

   (a) Taking the term $\beta_{ik}^{(13)}$ as your example, explain the constraints which are usually imposed on the terms in the log-linear model and how this leads to the number of free parameters (degrees of freedom) for each term.

   (b) Define what is meant by the term hierarchical log-linear model and draw a diagram showing all hierarchical log-linear models, which include all three main effects, for a three-way contingency table. Highlight in your diagram pairs of models which differ by a single term.

   (c) Taking two suitable models from your diagram as an example, explain in detail how model comparison works using the Akaike information criterion and the generalised likelihood ratio test. You need not explain how to compute the likelihood for these models.

   (d) Give an example of two models which would not be suitable for use in part(c) of the question and explain why.

4. The dataframe `horseshoe` contains the results of a study of nesting horseshoe crabs (J. Brockman, Ethology, 1996). Each female crab in the study had a male crab in her nest. The study investigated factors that affect how many other males, called satellites, the female has residing nearby. Variables thought to affect this include the female crab's colour, the condition of its spine, its weight, and its carapace width. The response variable for each female crab is the number of satellites.

`Co`	carapace colour (1: light medium; 2: medium; 3: dark medium; 4: dark)
`Sp`	condition of spine (1: both good; 2: one worn or broken; 3: both worn or broken)
`Wi`	width of carapace in cm
`We`	weight in kg
`Sa`	number of satellite male crabs

Consider the generalized linear model fitted in the R excerpt below.

(a) For each of the variables in the dataframe, state whether it is categorical or numerical.

(b) Why can we not use a linear model to explain the dependence of `Sa` on the other variables?

(c) For this model, write down expressions for:

- the linear predictor;
- the response function;
- the distributional assumption.

(d) Give a precise numerical interpretation of the estimated parameter for `We`.

(e) A new female crab is observed, with a dark carapace, carapace width 25.0cm, and weight 4.5kg. Use the model to predict the expected number of satellite male crabs for this female.

(f) Does it seem reasonable to treat the colour variable in this way?

```
> head(horseshoe)
  Co  Sp    Wi    We  Sa
1  2   3  28.3  3.05   8
2  3   3  22.5  1.55   0
3  1   1  26.0  2.30   9
4  3   3  24.8  2.10   0
5  3   3  26.0  2.60   4
6  2   3  23.8  2.10   0
> model = glm(Sa ~ Co + Wi + We, family = poisson, data = horseshoe)
> model$coefficients
(Intercept)          Co          Wi          We
-0.99307525 -0.16310297  0.05880436  0.34685788
```

5. Consider a logistic regression problem with predictor $x \in \mathbb{R}$ and response $y \in \{0, 1\}$. Let the data be $\{(x_i, y_i)\}_{i \in [1..n]}$. Suppose that $\max\{x_i : y_i = 0\} < \min\{x_i : y_i = 1\}$.

   (a) Prove that the maximum likelihood estimate of the linear predictor based on this data is not finite.

   (b) The resulting regression function exists but is not unique. Describe the nature of the function and its non-uniqueness.

   (c) Describe very briefly two circumstances in which the results of the previous two parts are a problem, and explain in each case what might be done to remedy the situation.

6. The data shown below give the survival times `S` of $n = 48$ animals, with four animals randomly allocated to each of the 12 possible combinations of 3 poisons and 4 antidotes. The experiment was part of an investigation to combat the effects of certain toxic agents.

poison											
I				II				III			
antidote				antidote				antidote			
A	B	C	D	A	B	C	D	A	B	C	D
0.31	0.82	0.43	0.45	0.36	0.92	0.44	0.56	0.22	0.30	0.23	0.30
0.45	1.10	0.45	0.71	0.29	0.61	0.35	1.02	0.21	0.37	0.25	0.36
0.46	0.88	0.63	0.66	0.40	0.49	0.31	0.71	0.18	0.38	0.24	0.31
0.43	0.72	0.76	0.62	0.23	1.24	0.40	0.38	0.23	0.29	0.22	0.33

A GLM was fitted and an analysis of deviance carried out, yielding the (edited) R code and output below, which you may use to answer the following questions.

(a) Complete the missing values for `W`, `X`, `Y`, and `Z`.

(b) Provide an estimate of the shape parameter of the response distribution.

(c) Test the model M_0, which only contains an intercept, against the full model M_2 using both `antidote` and `poison`, at the 5% level of significance.

(d) In this analysis, R has *not* treated these data as grouped, but as $n = 48$ individual observations. Explain how a grouped version of this data set would be obtained, how many observations it would contain, and write down explicitly the first three rows of the data frame containing the grouped data.

```
> fit<- glm(S~antidote+poison, data=animals, family=Gamma(link=log))
> anova(fit)
Analysis of Deviance Table
Model: Gamma, link: log
Response: S
Terms added sequentially (first to last)

         Df Deviance  Resid.Df  Resid.Dev
NULL                        47    11.5710
antidote  W        X         Y          Z
poison     2   5.0973        42     2.4036
> summary(fit)$dispersion
[1] 0.06127824
```

SECTION B

7. The negative binomial distribution, with parameters $\psi > 0$ and $p \in (0, 1]$, has probability function

$$\mathrm{P}[X = x; \psi, p] = \frac{\Gamma(\psi + x)}{\Gamma(\psi)\Gamma(x + 1)} p^{\psi}(1 - p)^{x} \quad \text{for } x = 0, 1, 2, \ldots$$

The R transcript, shown below and on the following page, shows the distribution being fitted to a dataset on number of days of absence for each student in a group of students attending a Los Angeles high school. The data are assumed to be a random sample from a negative binomial distribution.

(a) Write down the log-likelihood function for $\theta = (\psi, p)$ and show that it can be expressed in the form used to define the function `L` shown in the R transcript.

(b) Calculate 99% confidence intervals for both parameters. Explain how you can tell which parameter estimate is which from the R code.

(c) The mean of the distribution is $\mu = \psi(1 - p)/p$. Compute an approximate 95% confidence interval for μ.

(d) Determine the profile log-likelihood function for ψ.

(e) Suppose that the parameterisation (μ, σ) is of interest, where $\sigma^2 = \psi(1 - p)/p^2$.

(i) Show how to compute the estimated sampling variance matrix of the maximum likelihood estimate for this parameterisation. You should do any necessary mathematical calculations but need not do numerical calculations.

(ii) Write a short R function to evaluate the log-likelihood for this parameterisation, making use of the existing code where possible.

```
> L = function(theta) {
+   p = theta[1]
+   psi = theta[2]
+   n = length(absences)
+   constpart = n * ( psi * log(p) - lgamma(psi) )
+   datapart1 = sum(lgamma(psi+absences) - lgamma(absences+1))
+   datapart2 = sum(absences) * log(1-p)
+   return(constpart+datapart1+datapart2)
+ }
> negL = function(theta) {
+   return(-L(theta))
+ }
> optim(c(.5, 1), negL, method="BFGS", hessian=TRUE)
$par
[1] 0.1225473 0.9684335

$value
[1] 501.8582
```

```
$counts
function gradient
      48       14

$convergence
[1] 0

$message
NULL

$hessian
          [,1]       [,2]
[1,] 12274.565 -1362.7691
[2,] -1362.769   215.9963
```

8. (i) The full (saturated) log-linear model for a two-way contingency table having R rows and C columns is

$$\log p_{ij} = \eta_{ij} = \beta + \beta_i^{(1)} + \beta_j^{(2)} + \beta_{ij}^{(12)}$$

(a) Show that the two variables are independent if and only if there exist g_i $(i = 1, \ldots, R)$ and h_j $(j = 1, \ldots, C)$ such that $p_{ij} = g_i h_j$ for all i and j. Hence show that they are independent if and only if $\beta_{ij}^{(12)} = 0$ for all i and j. You may assume but should state the usual constraints that are applied to the terms in the model.

(b) Write down the log-likelihood (assuming random sampling) and use the method of Lagrange multipliers to show that the maximum likelihood estimates in the independence model satisfy $\hat{p}_i = y_i/n$ and $\hat{p}_j = y_j/n$.

(ii) The following table shows data for a three-way contingency table. Each variable has two levels.

	$k=1$		$k=2$	
	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	2	6	4	0
$i=2$	1	3	5	3

Carry out one full iteration of the iterative proportional fitting algorithm for the "no three-way interaction" model.

Is another iteration needed?

9. In an experiment on the number of animals that survived a treatment, the following results were observed:

	dead	alive	
not treated	30	25	$m_1 = 55$
treated	36	14	$m_2 = 50$

We can consider this as a data set which is grouped with respect to a binary covariate, with values (say) $z_1 = 0$ (not treated) and $z_2 = 1$ (treated), and with the response defined through group-wise survival rates: $y_1 = 25/55$ and $y_2 = 14/50$. We model this data through a binomial logit model, i.e

$$\pi(z) = \frac{\exp(\beta_1 + \beta_2 z)}{1 + \exp(\beta_1 + \beta_2 z)} ,$$

where $y_i|z_i \sim \text{Bin}(m_i, \pi(z_i))/m_i$ (the 'rescaled' binomial distribution).

(a) Verify through differentiation of the log-likelihood that the score function is given by:

$$S(\beta_1, \beta_2) = \sum_{i=1}^{2} m_i \begin{pmatrix} 1 \\ z_i \end{pmatrix} (y_i - \pi(z_i)) .$$

(b) Solve the score equation.

(c) Show that the expected Fisher information, evaluated at the ML estimate, is given by

$$F(\hat{\beta}_1, \hat{\beta}_2) = \begin{pmatrix} 23.72 & 10.08 \\ 10.08 & 10.08 \end{pmatrix} .$$

(d) Carry out a Wald-test of the null hypothesis $H_0 : \beta_2 = 0$ at the 5% level of significance.

(e) Compute the maximum likelihood estimate of β_1 under the constraint $\beta_2 = 0$. Hence carry out a likelihood ratio test of $H_0 : \beta_2 = 0$ at the 5% level of significance.

(f) Give an interpretation of your results.

(g) Explain why Wald tests can be less reliable than likelihood ratio tests.

10. (a) An exponential dispersion family (EDF) of probability distributions has probability density function of the following form:

$$P(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right] .$$

(i) Comment on the roles of the parameters θ and ϕ.

(ii) Prove that the mean $\mu = \mathrm{E}[Y|\theta, \phi]$ and variance $\mathrm{Var}[Y|\theta, \phi]$ of a member of an EDF are given by:

$$\begin{aligned} \mu &= b'(\theta) \\ \mathrm{Var}[Y|\theta, \phi] &= \phi\, b''(\theta) , \end{aligned}$$

where a prime indicates a derivative.

(iii) Why is b' usually invertible for almost all values of θ, and why is this important?

(b) The Gamma distribution has density

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

where $y \in \mathbb{R}_{\geq 0}$, and $\alpha, \beta \in \mathbb{R}_{>0}$.

(i) Show that the Gamma distribution is an EDF, identifying all the components of the model.

(ii) Exploiting properties of the EDF, calculate the mean of the Gamma distribution in terms of α and β.

(c) The Gamma distribution is used in a GLM. For data $\{(x_i, y_i)\}_{i\in[1..n]}$, let $\hat{\mu}_i$ be the estimated value of the expectation of Y at x_i, computed using the maximum likelihood estimate $\hat{\beta}$ of the GLM parameters.

(i) Explain the concept of the saturated version of the model, and derive the MLEs $\tilde{\mu}_i$ of $\mathrm{E}[Y|\theta, \phi]$ for each i.

(ii) Derive an expression for the deviance of the Gamma GLM.

(iii) Why does the deviance fail to be a complete replacement for RSS as a measure of model adequacy?

SECTION C

11. (i) The following data record the residue level of a pesticide in 10 apples randomly sampled from a large orchard:

0.10 0.12 0.16 0.20 0.22 0.24 0.26 0.32 0.46 0.66

The bootstrap will be used to make inference about the population median.

(a) Explain how, by using either dice or a uniform random number generator or R, you would take a bootstrap re-sample from the data.

(b) The following are the first two of 10000 bootstrap re-samples:

0.12 0.12 0.22 0.24 0.24 0.26 0.26 0.46 0.46 0.66

0.20 0.24 0.26 0.46 0.46 0.66 0.66 0.66 0.66 0.66

The first ten bootstrap statistics were:

0.25 0.56 0.25 0.22 0.22 0.20 0.29 0.22 0.24 0.23

How were these obtained? Be explicit in the case of the first two re-samples.

(c) The mean and standard deviation of the 10000 bootstrap statistics were respectively 0.235 and 0.048. The following table provides a number of percentiles:

Min.	1%	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%	Max.
0.10	0.14	0.16	0.16	0.18	0.21	0.23	0.25	0.29	0.32	0.36	0.39	0.66

Showing your working, calculate 95% bootstrap confidence intervals for the population median, using: (1) the *basic* method; (2) the normal approximation to the bootstrap sampling distribution.

(ii) Suppose that we want to make inference about the population mean from a sample of size 4 and that the sampled values turn out to be: $a - 2b$, $a - b$, $a + b$, $a + 2b$ for some real numbers a and b.

By enumerating the relevant re-samples from the non-parametric bootstrap, find the basic bootstrap confidence interval with approximately 96% nominal confidence; you may assume an effectively infinite resample size.

Compare the basic bootstrap confidence interval to the normal approximation bootstrap confidence interval for the same nominal confidence level and to the usual t-distribution based confidence interval for a population mean. Comment on the strengths and weaknesses of the three approaches. You may use the fact that $t_{3,.02} = 3.5$.

(iii) Suppose that an infinite population is discrete, i.e. it only has M distinct values: $x_1, \ldots, x_M$. Denote by p_i the proportion of the population which takes value x_i. Let x^* be the value of x_i with the smallest p_i and denote that p_i by p^*.

Suppose that $p^* \approx 0.01$. We take a sample of size n from the population. How large does n have to be so that x^* has probability 0.95 of appearing at least once in the sample?

Discuss the implication for applying the bootstrap to learning about a population parameter which is strongly dependent on rare values in the population. Give an example of such a parameter.