Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2021	MATH3361-WE01

Title: Topics in Statistics III

Time (for guidance only):	3 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: There is no restriction on the model of calculator which may be used.

Instructions to Candidates:	Credit will be given for your answers to all questions. All questions carry the same marks. Please start each question on a new page. Please write your CIS username at the top of each page. To receive credit, your answers must show your working and explain your reasoning.

Revision:	

Q1 **1.1** Consider a random sample $X_1,...,\ X_n$ of n independent and identically distributed observables. Assume a statistical model, $X_i \overset{\text{IID}}{\sim} f(\cdot|\theta)$, with unknown parameter $\theta \in \mathbb{R}$ and Fisher information $\mathcal{I}(\theta)$. Assume the statistical model satisfies the assumptions stated in Cramer's theorem for the asymptotic behavior of MLEs. Let θ_0 be the real value of the parameter θ in the statistical model. Let $\hat{\theta}_n$ be the MLE of parameter $\theta \in \mathbb{R}$. Let $g(\cdot)$ be a real valued function differentiable $m \geq 1$ times at θ_0 with $\frac{\mathrm{d}^m}{\mathrm{d}x^m} g(x)\Big|_{x=\theta_0} \neq 0$ but $\frac{\mathrm{d}^k}{\mathrm{d}x^k} g(x)\Big|_{x=\theta_0} = 0$ for $k < m$. Show that

$$m! n^{m/2} \frac{g(\hat{\theta}_n) - g(\theta_0)}{\frac{\mathrm{d}^m}{\mathrm{d}x^m} g(x)\Big|_{x=\theta_0} [\mathcal{I}(\theta_0)]^{-m/2}} \xrightarrow{\text{D}} z^m$$

where $z \sim \text{N}(0, 1)$. Explain your working and justify your steps.

1.2 Consider the following data on worldwide airline fatalities from 1976 to 1985.

year (t)	passenger deaths y	passenger miles m (100 Million)
1976	734	3863
1977	516	4300
1978	754	5027
1979	877	5481
1980	814	5814
1981	362	6033
1982	764	5877
1983	809	6223
1984	223	7433
1985	1066	7107

The death counts can be described through a Poisson distribution $y \sim \text{Poi}(\mu)$, where the passenger miles m play the role of an *exposure variable*. That is, the quantity being modelled (as a function of time) is the expected rate $\lambda = \mu/m$ of passenger deaths per 100 million miles, with

$$\ln(\lambda) = \beta_1 + \beta_2 t.$$

Below is the (shortened) R output from the model fitted to these data (where `pdeaths` = y, `pmiles` = m, and `year` = t).

(i) Carry out a goodness-of-fit test for the Poisson model and conclude whether or not the Poisson model is appropriate for this situation.

(ii) Calculate two estimates of the dispersion, one using a quick method and another using a model-based method. Do they both support the conclusion reached in part (i)? Note that the variance function of the Poisson distribution is $v(\mu) = \mu$.

```
> summary(glm(pdeaths ~ year, offset = log(pmiles),
family = poisson(link=log), data = airline))
...
Coefficients:
```

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) 117.767873   8.420250   13.99   <2e-16 ***
year         -0.060522   0.004252  -14.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1253.6  on 9  degrees of freedom
Residual deviance: 1051.4  on 8  degrees of freedom
...
```

Q2 You work in a company publishing academic textbooks. Your company collaborates with two printing companies, A and B, to which orders are assigned randomly. Your books are available in two bindings: paperback and hardcover. Your company has recently received several complaints about the quality of the books, and your CEO has asked you to perform statistical analysis with the aim of continuing to use the services of the printing company that performs better. You have collected a random sample of orders, and classified them according to the following three variables: the printing company (X) with levels (A or B), the binding type (Z) with levels (Paperback or Hardcover), and the reported complaint (Y) with levels (No or Yes). The data-set is presented in Table 1.

		Reported complaint (Y)	
Binding type (Z)	Printing company (X)	No	Yes
Hardcover	A	500	15
	B	6000	150
Paperback	A	5000	25
	B	2000	4

Table 1: Dataset

2.1 Calculate the marginal XY-contingency table of the observed counts. Calculate a 95% confidence interval for the marginal odds ratio of the Printing company and Reported complaint, and based on this, infer whether the Printing company and Reported complaint are dependent or not. What does the estimated odds ratio of the Printing company and Reported complaint tell us about the relation between Printing company and Reported complaint?

2.2 We are interested in checking the hypothesis that the Reported complaint and Binding type are independent of the Printing company against the hypothesis that the Printing company and Reported complaint are conditionally independent given the Binding type. State the log-linear model equations for the two associations under consideration. Perform a statistical test at significance level 5% in order to test the hypothesis. Show and explain your workings.

2.3 Compute the conditional odds ratio of the Printing company and Reported complaint at each level of Binding type. What do these conditional odds say about the association between the classification variables involved?

2.4 Calculate the marginal odds ratio of the Printing company and Binding type. Use an appropriate inferential tool based on the odds ratio statistic to infer about the association between the Printing company and Binding type.

2.5 Compare the inferential results from parts (**2.1**) and (**2.3**). Explain why this behaviour might occur in our data analysis. Which Printing company will you suggest that your CEO continue to use?

Q3 **3.1** Consider a random sample of unseen observations $Y_1, ..., Y_n$ independently and identically distributed according to a distribution admitting density

$$f_Y(y) = \begin{cases} \left(\frac{\lambda}{2\pi y^3}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\phi y + \frac{\lambda}{y} - 2\sqrt{\phi\lambda}\right]\right) & , y > 0 \\ 0 & , y \leq 0 \end{cases}$$

with unknown parameter $\theta = (\lambda, \phi)$ where $\lambda > 0$ and $\phi > 0$. By collecting a sample $y_1, ..., y_n$, we are interested in making inference about the parameter λ.

(i) By profiling out ϕ , show that the MLE are given by

$$\hat{\phi}_\lambda = \frac{\lambda}{\bar{y}^2}, \quad \text{and} \quad \hat{\lambda} = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i} - \frac{1}{\bar{y}}\right]^{-1} \quad \text{where} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(ii) Derive an appropriate confidence interval based on the Wald statistic for the unknown parameter λ at significance level a. Derive appropriate rejection areas for the hypothesis test with hypotheses

$$\mathrm{H}_0 : \lambda = \lambda_* \ \text{ vs } \ \mathrm{H}_1 : \lambda \neq \lambda_*$$

based on a suitable Score statistic and on the likelihood ratio statistic at significance level a.

3.2 Let $\{Y_j\}_{j=1}^n$ be independent identically distributed random variables with distribution $f(\cdot|\theta)$ depending on the unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$. Let $\ell_n(\theta)$ denote the log-likelihood function. Let $\tilde{\theta}_n$ denote a preliminary estimate of θ such that $\tilde{\theta}_n = \theta + O_p(n^{-1/2})$. Let T_n denote a one-step estimator according to Newton's iteration:

$$T_n = \tilde{\theta}_n - \left[\ddot{\ell}_n(\tilde{\theta}_n)\right]^{-1} \dot{\ell}_n(\tilde{\theta}_n)$$

Assume that the first four derivatives of $\ell_n(\theta)$ exist and they are bounded in probability by $O_p(n^{-2})$, for instance $\dot{\ell}_n(\theta) = O_p(n^{-2})$, $\ddot{\ell}_n(\theta) = O_p(n^{-2})$. Further assumptions about the sampling density $f(\cdot|\theta)$, such as those in Cramer's theorem for MLE asymptotics are not necessarily assumed.

(i) Let $\mathcal{I}(\theta)$ denotes the Fisher information matrix. Prove that

$$\frac{1}{\sqrt{n}}\dot{\ell}_n(\tilde{\theta}_n) = \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_n) - \mathcal{I}(\theta)\sqrt{n}(\tilde{\theta}_n - \theta) + O_p\left(n^{-\frac{1}{2}}\right)$$

and

$$-\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) = \mathcal{I}(\theta) + O_p\left(n^{-\frac{1}{2}}\right)$$

(ii) Prove that

$$\sqrt{n}\,(T_n - \theta) = [\mathcal{I}(\theta)]^{-1}\,\frac{1}{\sqrt{n}}\dot{\ell}_n(\theta) + O_p\left(n^{-\frac{1}{2}}\right) \ ,$$

find the asymptotic distribution of $\sqrt{n}\,(T_n - \theta)$ along with the type of convergence, and argue whether T_n is asymptotically efficient.

Hint: Use part **3.2**.(i).

Q4 An investigation is to be undertaken into how the success or otherwise of penalties taken in English Premier League football matches depends on the position of the penalty taker (forward or midfield). Note that to take a penalty means that a player takes a single shot on the goal while it is defended only by the opposing team's goalkeeper. Data are gathered on penalties taken by forward and midfield players in a number of matches. The dataset has two variables: one records the position of the penalty taker (either "forward" or "midfield") and the other records the outcome of the penalty (1 if the penalty is scored and 0 if it is not).

First, suppose that a linear model is fitted to the data, with the penalty outcome as the response variable Y. Output from fitting the linear model is shown in the table below.

Parameter	Estimate	Estimated Standard Error
Intercept	0.850	0.06423
Midfield	-0.1000	0.08543

4.1 According to this linear model, what is the estimated probability that a forward scores a penalty? What is the corresponding estimated probability for a midfielder?

Now, a logistic model is fitted to the same data. Output from fitting the logistic model is shown in the table below.

Parameter	Estimate	Estimated Standard Error
Intercept	1.7346	0.4428
Midfield	-0.6360	0.5465

4.2 According to this logistic model, what is the estimated probability that a forward scores a penalty? What is the corresponding estimated probability for a midfielder?

4.3 How do the estimates from the two models relate to each other? Explore this by deriving in detail formulae for the estimated probabilities for the two models.

4.4 According to the results of the logistic model in the above table, does a player's position affect the probability of a successful penalty? Write your full hypothesis testing problem and use a significance level of 0.05.

4.5 One potentially interesting variable missing from the dataset is the identity of the player taking the penalty. What would be the benefit of having this information? What might be a disadvantage? To what extent does inclusion of player identity affect the conclusion of part **4.4**?

Q5 Consider a generalized linear model with inverse Gaussian distributed response, i.e. with probability density function:

$$P(y \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{-\left(\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)\right\}$$

where $y \in \mathbb{R}_{\geq 0}$ and $\mu, \lambda \in \mathbb{R}_{>0}$; and also suppose $E[y|x, \beta] = h(\beta^T x)$ for some response function h.

5.1 The inverse Gaussian distribution forms an exponential dispersion family with dispersion parameter $\phi = 2/\lambda$. Identify the natural parameter θ and the log normalizer $b(\theta)$, and hence derive the canonical link for this model. Why is this choice of link function problematic in general?

From now on, use the log link function.

5.2 Given data $\{(x_i, y_i)\}_{i\in[1..n]}$, write down the log likelihood for the model and derive the score function.

5.3 Derive the observed Fisher Information $F_{\text{obs}}(\boldsymbol{\beta})$ and the (expected) Fisher Information $F(\boldsymbol{\beta})$.

5.4 We are given measurements of peak ground acceleration `a` (in units of g) resulting from 10 different seismic events at different distances `d` (in tens of km) from the observation station:

`d`	1.20	12.3	1.96	9.10	3.29	0.38	2.20	1.22	2.90	4.92
`a`	0.359	0.003	0.200	0.039	0.064	0.640	0.150	0.097	0.039	0.017

In order to try to estimate the attenuating effect of distance on acceleration, a generalized linear model of the type described above is fitted to these data, with linear predictor $\eta = \beta_1 + \beta_2 \texttt{d}$. The estimates for β were found to be $\hat{\beta}_1 = 3.172$ and $\hat{\beta}_2 = -0.729$.

(i) How far must one travel in order to halve the expected peak acceleration?

(ii) Compute the expected Fisher Information matrix.

(iii) Provide an approximate test of $H_0 : \beta_2 = 0$, and comment whether or not it is significant.

(iv) Give the expected value of `a` and an approximate 95% confidence interval for this expected value when `d` = 0.

(v) Would you trust this expected value in practice? Why?