McGILL UNIVERSITY
FACULTY OF MEDICINE

Department of Epidemiology, Biostatistics and Occupational Health

PhD Program in Biostatistics: BIOS 700 Part A Comprehensive Exam

Theory Paper

Date: May 13, 2014 Time: 13:00 - 17:00

INSTRUCTIONS

Answer two questions from Section BS: one of BS1 and BS2, and one of BS3 and BS4. If you answer more than two questions, ONLY THE FIRST FROM EACH PAIR of questions (BS1/BS2 and BS3/BS4) will be marked.

Answer four questions from Section S. If you answer more than four questions, only the FIRST FOUR questions will be marked.

Questions	Marks
BS1	
BS2	
BS3	
BS4	
S1	
S2	
S3	
S4	
S5	
S6	

This exam comprises this cover page, and 10 pages of questions.

(§BS)

Answer only two questions in this section: (BS1 or BS2) and (BS3 or BS4)

BS1. 20 MARKS

Unless specifically asked to, do NOT complete the algebra.

(a) State the 'redistribution to the right' procedure that Efron used as an alternative derivation of the Kaplan-Meier survival function estimator. If it helps, use small numerical example, but there is no need to carry out the calculations to the very end. [3 MARKS]

(b) The 'exponential formula' $S[t] = \exp[-\int_0^t h(u)du]$ is often used to convert a hazard function to a t-year survival probability.

The Nelson-Aalen estimator of the survival probability $S[t]$ involves the number of deaths d_j and the number at risk n_j in each risk set up to time t.

Show that if one uses this 'exponential formula' one arrives at the Nelson-Aalen estimator

$$\widehat{S[t]}_{NA} = \exp[-\sum d_j/n_j],$$

with the summation over the risk sets up to time t. [5 MARKS]

(c) Suppose, for each of n unrelated children, we knew the father's age (a) at the time the child was conceived, and the number (y) of new mutations he passed on to his child (by 'new', we mean that mutation occurred between the time the father was born and the time the child was conceived)

i. Assume that the *de novo* mutation rate λ (so that the expected number of mutations that occur in the short interval δt is $\lambda \times \delta t$) is independent of (i.e., constant over) a man's age (life),

$$\text{Mutation Rate at Age } a = \lambda, \ \forall a,$$

and that the mutations found in his children are all transmitted from him, and that none are inherited from the children's mother.
Derive an ML estimator for λ, stating any assumptions made.
[4 MARKS]

ii. Write down GLM code that would yield $\widehat{\lambda_{ML}}$. [2 MARKS]

iii. Suppose you wished to obtain ML estimates for the two parameters λ_0 and β in the age-dependent mutation rate model:

$$\text{Mutation Rate at Age } a = \lambda_0 + \beta \times a, \qquad (\textit{additive rate model})$$

Derive $E(y|a)$ and modify the GLM code so as to fit the two parameters. [3 MARKS]

iv. Describe how you would fit the model

Mutation Rate at Age $a = \lambda_0 \times \exp[\beta \times a]$, (*multiplicative model*).

You don't need to finish all the steps: just describe the broad approach. [3 MARKS]

BS2. 20 MARKS

Unless specifically asked to, do NOT complete the algebra.

(a) This is the `summary` of the delays (measured in days) for some $n = 250{,}000$ scheduled medical procedures:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D.
0	0	0	10.8	7	1627	32.5

The mean of 10.8 days was accompanied by a 95% CI with a margin of error of $1.96 \times 32.5/\sqrt{250,000}$ days.

i. Comment on the appropriateness/inappropriateness of the 1.96 in the margin of error. [3 MARKS]

ii. Suppose you were doing a similar study, but your budget was limited to $n = 400$ observations. How would you calculate margins of error to accompany the mean you obtained? [4 MARKS]

(b) In surveys in which people are asked to prospectively record their purchasing behaviour in a given product category over a specified time period, some people will forget (or intentionally neglect) to record all of their actual purchases. Consequently, the recorded purchase counts under-estimate the actual level of purchasing behaviour.

The recorded count for a person might be modelled by denoting the actual and recorded number of purchases of that person as random variables Y and Y', respectively, so that

$$Y' = \sum_{j=1}^{j=Y} I_j,$$

where I_j is an indicator variable taking on the value 1 if the person's jth purchase is recorded, 0 otherwise. Assume that the actual number of purchases is Poisson distributed with mean parameter μ and that the I_j's are i.i.d. Bernoulli with parameter π. These assumptions lead to what some authors call a 'Poisson-stopped' sum of Bernoulli random variables.

i. How would you derive the expectation and variance of Y' ? It is not necessary to finish the derivation: it is enough to show how one starts, and how the algebra ends. [3 MARKS]

ii. How would you show that Y' is indeed Poisson distributed with mean parameter $\mu\pi$. Again, it is not necessary to finish the derivation: it is enough to show how one starts, and how the algebra ends. [3 MARKS]

iii. Suppose now that the distribution of μ across individuals is gamma with shape parameter θ_1 and rate parameter θ_2 (or scale parameter $1/\theta_2$.) What is the name of the marginal distribution of Y? [1 MARKS]

iv. Outline how would you formally prove that it has this distribution. [2 MARKS]

v. Suppose you had the t*rue* counts $\{y_1, \ldots, y_n\}$ for a random sample of n persons. Indicate very briefly how you would use them to fit θ_1 and θ_2. [2 MARKS]

vi. Suppose now that the distribution of π across individuals is beta with parameters θ_3 and θ_4, and that an individual's recording probability (p) is independent of his/her purchase rate (λ). Indicate very briefly a method you might you use just the *recorded* counts $\{y'_1, \ldots, y'_n\}$ for a random sample of n persons, to fit θ_1 to θ_4. State any additional assumptions you would need to make. [2 MARKS]

BS3. 20 MARKS

Please DO complete the algebra as well as the arithmetic unless specifically asked not to.

Suppose that we are interested in the association between ischaemic heart disease (IHD) incidence and daily energy intake. Let us denote the index level of exposure (energy intake less than 2750 kcals per day) as $Z = 1$ and the reference level of exposure (energy intake at least 2750 kcals per day) as $Z = 0$. Further, let X be an age band indicator taking values $X = 0$ (40 - 49), $X = 1$ (50 - 59) and $X = 2$ (60 - 69). The observed data tabulated by the age band and exposure are given in Table 1.

(a) We are in particular interested in the difference of the expected numbers of IHD events in randomly drawn 4626.4 person-years of population-time, had this population-time been exposed, versus unexposed. Formulate the parameter of interest (a marginal causal contrast) in terms of potential outcome variables. Remember to explain your notation. [2 MARKS]

(b) Let D denote the observed number of IHD events in randomly drawn 4626.4 person-years of population-time in age band X and exposure category Z. Show that

$$E\left[\frac{ZD}{P(Z=1 \mid X)}\right] - E\left[\frac{(1-Z)D}{P(Z=0 \mid X)}\right]$$

is equivalent to the causal parameter in (a). If showing this requires further definitions/assumptions, please also state these. [5 MARKS]

(c) Use the result in (b) to motivate a model-based direct standardization estimator for the causal parameter in (a). [5 MARKS]

(d) Based on the data in Table 1 and the model output in Table 2, calculate a model-based direct standardization estimate for the causal parameter in (a). [5 MARKS]

(e) How would you calculate a model-free direct standardization estimate for the causal parameter in (a)? (Explain but do not calculate a numerical value.) [3 MARKS]

BS4. 20 MARKS

Please DO complete the algebra as well as the arithmetic unless specifically asked not to.

Suppose that we are interested in the association between ischaemic heart disease (IHD) incidence and daily energy intake. Let us denote the index level of exposure (energy intake less than 2750 kcals per day) as $Z = 1$ and the reference level of exposure (energy intake at least 2750 kcals per day) as $Z = 0$. Further, let X be an age band indicator taking values $X = 0$ (40 - 49), $X = 1$ (50 - 59) and $X = 2$ (60 - 69). The observed data tabulated by the age band and exposure are given in Table 1.

(a) Specify the regression model corresponding to the output in Table 2. Remember to explain your notation. [3 MARKS]

(b) Show (algebraically) how the regression coefficient for the exposure can be interpreted in terms of a rate ratio, and explain why such a model is a special case of proportional hazards model. [3 MARKS]

(c) Write the likelihood function resulting from the model specification in (a) and the data in Table 1. [4 MARKS]

(d) From the likelihood function obtained in (c), derive a Cox partial likelihood form profile likelihood function for the regression parameters which eliminates the baseline rate parameter. (Do this algebraically, without substituting in the numbers in Table 1) [6 MARKS]

(e) Calculate the model-based 10-year IHD risk difference between exposed and unexposed 40-year old individuals. How does this compare to the corresponding model-free 10-year IHD risk difference? (If there is a discrepancy, explain the reason for this.) [4 MARKS]

Table 1: Person-years and numbers IHD events by age band X and exposure status Z.

		Person-years	Observed IHD events
$X = 0:$	$Z = 0$	607.9	4
	$Z = 1$	311.9	2
		Person-years	Observed IHD events
$X = 1:$	$Z = 0$	1272.1	5
	$Z = 1$	878.1	12
		Person-years	Observed IHD events
$X = 2:$	$Z = 0$	888.9	8
	$Z = 1$	667.5	14

Table 2: A regression model fitted to the data in Table 1.

```
Call:
glm(formula = d ~ z + as.factor(x) + offset(log(y)),
    family = poisson(link = "log"))

Deviance Residuals:
       1         2         3         4         5         6
 0.73940  -0.58410   0.04255  -0.77385   0.42800  -0.03191

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.4177     0.4421 -12.256  < 2e-16 ***
z               0.8697     0.3080   2.823  0.00476 **
as.factor(x)1   0.1290     0.4754   0.271  0.78609
as.factor(x)2   0.6920     0.4614   1.500  0.13366
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14.5780  on 5  degrees of freedom
Residual deviance:  1.6727  on 2  degrees of freedom
AIC: 31.796

Number of Fisher Scoring iterations: 4
```

(§S.) Answer only 4 questions out of S1-S6

S1. 20 MARKS

Consider a Dirichlet distributed random vector (X_1, X_2, X_3) with parameters $\alpha_1, \alpha_2, \alpha_3 > 0$, that is, $X_3 = 1 - X_1 - X_2$ and the density of (X_1, X_2) is

$$f(x_1, x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{\alpha_3 - 1}$$

for all $x_1, x_2 > 0$ such that $x_1 + x_2 < 1$.

(a) What can you say about the density of (X_1, X_2, X_3)? **(3 marks)**

(b) Determine the marginal distributions of X_i, $i = 1, \ldots, 3$. **(6 marks)**

(c) Compute the correlation between X_1 and $X_2 + X_3$. *Justify every step you make.*

(5 marks)

(d) Suppose that $Y_1 \sim \text{Beta}(\alpha_1, \alpha_2 + \alpha_3)$ and $Y_2 \sim \text{Beta}(\alpha_2, \alpha_3)$ are independent. Prove that

$$(X_1, X_2, X_3) \stackrel{d}{=} (Y_1, Y_2(1 - Y_1), (1 - Y_1)(1 - Y_2))$$

where $\stackrel{d}{=}$ denotes equality in distribution. *Hint: show first that* $(X_1, X_2) \stackrel{d}{=} (Y_1, Y_2(1 - Y_1))$. **(6 marks)**

S2. 20 MARKS

Consider the inverse Gaussian distribution with parameters $\lambda > 0$ and $\mu > 0$. Its density is given by

$$f(x;\lambda,\mu) = \frac{\sqrt{\lambda}}{\sqrt{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0.$$

(a) Show that the inverse Gaussian family of distributions is an exponential family. Identify the canonical parameters and determine the canonical parameter space.

(7 marks)

(b) Suppose that X is an inverse Gaussian random variable. Compute the correlation between X and $1/X$. **(7 marks)**

(c) Show that the correlation of any two random variables Y and Z with finite second moments satisfies $|\text{corr}(Y,Z)| \leq 1$. Can the bound $|\text{corr}(Y,Z)| = 1$ be attained for any pair (Y,Z) of random variables with given marginal distributions? Justify your answer. **(6 marks)**

S3. 20 MARKS

Suppose that $\alpha, \beta > 0$ and $(X_1, P_1), \ldots, (X_k, P_k)$ are independent random vectors such that

$$\begin{aligned} X_i|P_i &\sim \text{Binomial}(n_i, P_i), \quad i = 1, \ldots, k, \\ P_i &\sim \text{Beta}(\alpha, \beta). \end{aligned}$$

Denote the total number of successes by $Y = \sum_{i=1}^{k} X_i$.

(a) Compute the expectation and variance of Y. **(6 marks)**

(b) Determine the distribution of Y when $n_1 = \cdots = n_k = 1$. **(7 marks)**

(c) Suppose that W and Z are random variables with finite expectations. Determine a function h such that $W - h(Z)$ is orthogonal to $g(Z)$, viz.

$$\text{E}\Big[\big\{W - h(Z)\big\}g(Z)\Big] = 0,$$

for any measurable function g such that $\text{E}\{g(Z)\}$ is finite. *Show your work and justify every step you make.* **(7 marks)**

S4. 20 MARKS

Find a nontrivial set of sufficient statistics in each of the following cases:

(a) Random variables $X_{jk}(j = 1, \cdots, m \ ; \ k = 1, \cdots r)$ have the form $X_{jk} = \mu + \eta_j + \varepsilon_{jk}$, where the η_j's and the ε_{jk}'s are independently normally distributed with zero means and variances respectively σ_b^2 and σ_w^2. The unknown parameters are thus $(\mu, \sigma_b^2, \sigma_w^2)$. **(10 marks)**

(b) Independent binary random variables $Y_1, \cdots, Y_n$ are such that the probability of the value one depends on an explanatory variable x, which takes corresponding values $x_1, \cdots, x_n$, through the model

$$\log\left[\frac{P(Y_j = 1)}{P(Y_j = 0)}\right] = \gamma + \beta x,$$

where γ and β are scalar-valued constants. **(10 marks)**

S5. 20 MARKS

Let $X_i \overset{iid}{\sim} N(\theta, 1)$, $i = 1, 2, \cdots, n$. Consider the sequence

$$\delta_n = \begin{cases} \overline{X}_n, & \text{if } |\overline{X}_n| \geq 1/n^{1/4}, \\ a\overline{X}_n, & \text{if } |\overline{X}_n| \leq 1/n^{1/4}. \end{cases}$$

Show that $\sqrt{n}(\delta_n - \theta) \xrightarrow{\mathcal{L}} N(0, \nu(\theta))$, where $\nu(\theta) = 1$ if $\theta \neq 0$ and $\nu(\theta) = a^2$ if $\theta = 0$. Is $\nu(\theta)$ greater than or equal to the information bound? (**Hint:** condition on $|\overline{X}_n|$).

(20 marks)

S6. 20 MARKS

If we wish to study the distribution of X, the number of albino children (or children with a rare anomaly) in families with proneness to produce such children, a convenient sampling method is first to discover an albino child and through it obtain the albino count X^w of the family to which it belongs. If the probability of detecting an albino is β, then the probability that a family with k albinos is recorded is $w(k) = 1 - (1 - \beta)^k$, assuming the usual independence of Bernoulli trials. In such a case

$$p_{X^w}(k) = P(X^w = k) = \frac{w(k)P(X = k)}{\mathbb{E}[w(X)]}, \quad k = 0, 1, 2, \cdots$$

(a) Suppose X has the *Pascal Distribution*, that is

$$P(X = k) = \frac{\alpha^k}{(1 + \alpha)^{k+1}}, \quad k = 0, 1, 2, \cdots$$

Find $\mathbb{E}(X)$ and show that

$$\lim_{\beta \to 0} \frac{w(k)}{\mathbb{E}[w(X)]} = \frac{k}{\mathbb{E}(X)}.$$

State clearly the assumptions you need to establish this result.

(7 marks)

(b) Suppose β is small enough, such that the result of Part (a) is applicable. Is this probability distribution a member of Exponential family? Let $X_1^w, \cdots, X_n^w$ be a sample of size from p_{X^w}. Find a complete sufficient statistic for α.

(7 marks)

(c) Using the asymptotic distribution of α find a 95% confidence interval for α.

(6 marks)