UNIVERSITY OF SWAZILAND

EXAMINATION PAPER 2009

TITLE OF PAPER	: TOPICS IN STATISTICS (CATEGORICAL DATA ANALYSIS AND GENERALIZED LINEAR MODELS)
COURSE CODE	: ST 405
TIME ALLOWED	: TWO (2) HOURS
REQUIREMENTS	: CALCULATOR AND STATISTICAL TABLES
INSTRUCTIONS	: ANSWER ANY FIVE QUESTIONS

Question 1

a) Prove that for a 2 × 2 contingency table, the chi-square test for independence is given by

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

(12 Marks)

b) Suppose that independent random variables $Y_1, Y_2, \ldots, Y_n$ follow binomial distributions, that is

$$Y_i \sim \mathrm{B}(m_i, \pi_i),\ i = 1, 2, \ldots, n, \quad \text{where } P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}.$$

Show that the binomial distribution is a member of the exponential family and that natural canonical link function for this distribution in the context of generalised linear models is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

(8 Marks)

Question 2

a) A sports equipment company has commissioned an advertising agency to develop an advertising campaign for one of its new products. They can choose between two particular television commercials, *A* and *B*. To aid them in their decision, an experiment is performed in which 200 volunteers are randomly assigned to view one of the two commercials, 100 being assigned to each. After seeing the commercial, each volunteer is asked to state whether they would consider buying the product, with the following results.

		Commercial	
		A	*B*
Purchase product	*No*	70	80
	Yes	30	20

Apply a chi-squared test to these data and comment on your results. What recommendations, if any, would you make to the sports manufacturer concerning the choice of commercial for the proposed advertising campaign?

(15 Marks)

b) In an industry the maintenance department wants to investigate the number of repairs to be made to 4 different makes of compressors, which are used in the 3 areas (North, Centre and South). The data on compressor failures are as follows:

Compressor	North	Centre	South
1	17	17	12
2	11	9	13
3	11	8	19
4	14	7	28

They engage a consultant to analyse the data and make a recommendation. What would be your recommendation to the maintenance department?

(20 Marks)

Question 3

An art gallery is due to celebrate its 50th anniversary in 2005. As part of its celebrations, it wishes to commission a new sculpture to be displayed in the gallery. To find a suitable sculpture, it decided to run a competition in which it invited local artists to submit designs. A panel of experts selected a short-list of three designs for the gallery to choose from. To assist in the final decision, the gallery conducted a survey in which a random sample of local adults were sent copies of the three designs and asked to indicate their preference. The replies received from male and female adults are given in the following table.

	Preferred design		
	A	*B*	C
Males	129	24	47
Females	126	44	55

Carry out a suitable analysis to test whether or not the preference of design is the same for males and females.

(20 Marks)

Question 4

Given the following table with a set of models with values of G^2 (Likelihood Ratio Criterion) and *p*-value which relate to;

- Defendant's race **D** : *W (White), B (Black)* = Z variable
- Victim's race **V** : *W (White), B (Black)* = Y variable
- Death Penalty **P** : *Y (yes), N (No)* = X variable

Model	G^2	*p*-value
(D,V,P)	137.9	0.001
(DV,P)	8.1	0.04
(VP,DV)	1.9	0.39
(DP,VP,DV)	0.70	0.40
(DVP)	0	---

a) Select two models that provide the best fit for the data and state the reasons for your choice. (2 Marks)

b) From the two models select the best model and derive its log-linear model and give reasons for selecting it. (6 Marks)

c) If n_{111}=19, n_{112}=0, n_{121}=11, n_{122}=6, n_{211}=132, n_{212}=9, n_{221}=52, and n_{222}=97, for the best model chosen in b) calculate $\hat{u}$, $\hat{u}_{D(1)}$ and $\hat{u}_{DV(11)}$.

(12 Marks)

Question 5

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases.

A research team from the U.S. measured all wells in an area of Araizahar upazila and labelled them with their arsenic level as well as a characterization as "safe" or "unsafe", depending on whether the arsenic level was above or below the national standard of 0.5 in units of hundreds of micrograms per litre.

People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. The amount of water needed for drinking is low enough that adding users to a well would not exhaust its capacity. The surface water in this area is contaminated, hence the desire to use deep wells.

A few years later the researchers returned to see who had switched wells and found that 57.5% of the 3020 households with unsafe wells had switched. The team performed a series of analyses to understand the factors predictive of well switching among users of unsafe wells.
Variables:
distnear = the distance to the nearest safe well
ed = years of education
as = arsenic levels (ug/L)
logas = log-arsenic
edcXdistnc = (ed-med)*(distnear-mdistn)
med = mean of ed
mdist = mean of distnear

Model 1

```
                                                  Number of obs   =      3020
                                                  LR chi2(3)      =    239.95
                                                  Prob > chi2     =    0.0000
Log likelihood =  -1939.077

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 distnearest |  -.0097893   .0010616    -9.22   0.000    -.0118699   -.0077087
       logas |    .888925    .068873    12.91   0.000     .7539365    1.023913
          ed |   .0431016   .0096435     4.47   0.000     .0242007    .0620024
       _cons |  -3.776544   .3315441   -11.39   0.000    -4.426358   -3.126729
------------------------------------------------------------------------------
```

Model 2

```
                                                  Number of obs   =      3020
                                                  LR chi2(4)      =    253.48
```