UNIVERSITY OF SWAZILAND

MAIN EXAMINATION PAPER 2016

TITLE OF PAPER	: TOPICS IN STATISTICS (STATISTICAL MODELLING)
COURSE CODE	: ST 405
TIME ALLOWED	: THREE (3) HOURS
REQUIREMENTS	: CALCULATOR AND STATISTICAL TABLES
INSTRUCTIONS	: ANSWER ANY FIVE QUESTIONS

Question 1

a) Consider the following data from a women's health study (MI is myocardial infarction, i.e. heart attack).

	MI	
Oral contraceptives	**Yes**	**No**
Used	**23**	**34**
Never Used	**35**	**132**

(i) Construct a 95% confidence interval for the population odds ratio.

(ii) Suppose that the answer to part (a) is (1.3, 4.9). Does it seem plausible that the variables are independent? Explain.

b) For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4.

(i) What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males"?

(ii) When would the quoted interpretation be approximately correct? Why?

(iii) The odds of survival for females equalled 2.9. For each gender, find the proportion who survived.

(4+4+4+4+4+4 Marks)

Question 2

The following table and subsequent analysis are based on sample survey data on the usage of alcohol, cigarettes and marijuana among high school students;

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

R code

```
A<-c(1,1,1,1,0,0,0,0); ## 1--Alcohol use 0--otherwise
C<-c(1,1,0,0,1,1,0,0); ## 1---Cigarette use 0---otherwise
M<-c(1,0,1,0,1,0,1,0); ## 1-Marijuana use 0-otherwise
count<-c(911,538,44,456,3,43,2,279);
 AC<-A*C; AM<-A*M; CM<-C*M; ACM<-A*C*M;

##Model (AM,CM,AC) fit
drug.log<-glm(count~A+C+M+AM+CM+AC,family=poisson(link="log"))
summary(drug.log)

## output
 Call: glm(formula = count ~ A + C + M + AM + CM + AC, family =
poisson(link = "log"))  Coefficients:
            Estimate  Std. Error   z value            Pr(>|z|)
(Intercept)  5.63342     0.05970   94.361              < 2e-16***
A            0.48772     0.07577    6.437             1.22e-10 ***
C           -1.88667     0.16270  -11.596              < 2e-16 ***
M           -5.30904     0.47520  -11.172              < 2e-16***
AM           2.98601     0.46468    6.426            <1.31e-10***
CM           2.84789     0.16384   17.382              < 2e-16 ***
AC           2.05453     0.17406   11.803              < 2e-16 ***
 Null deviance: 2851.46098,   Residual deviance:    0.37399
##Estimated covariance matrix between AM and CM

          AM           CM
AM  0.215925578 -0.004968391
CM -0.004968391 0.026843349
```

Let X; Y and Z denote the variables Alcohol, Cigarette and Marijuana use respectively.

a) Write down the loglinear regression model and identify the associated estimates.
(5 Marks)

b) Compute the estimated odds ratio between any two variables of Alcohol, Cigarette, and Marijuana use controlling for the third variable.

(5 Marks)

c) Construct the 95% confidence interval for the true odds ratio between Alcohol and Cigarette use controlling for Marijuana use.

(5 Marks)

d) Test if the true odds ratio between Alcohol and Marijuana use controlling for Cigarette use equals the true odds ratio between Cigarette and Marijuana use controlling for Alcohol use at $\alpha = 5\%$.

(8 Marks)

Question 3

a) For each of the following densities for a random variable Y, show that Y or some transformation of Y has an exponential family distribution. Derive the mean and variance of the exponential family distributed quantity in each case using the mean and variance formulas that hold in general within the exponential family distribution.

(i) $f(y;\mu,\lambda) = (2\pi y^3/\lambda)^{-1/2} exp\left\{-\lambda/_{2u^2}(y-\mu)^2/y\right\}, \qquad y,\lambda,\mu > 0.$

(ii) $f(y;\theta) = \theta a^{\theta}/y^{(\theta+1)}, \qquad y > a, \theta > 0, a > 0.$

(20 Marks)

Question 4

For a classical linear model $= \boldsymbol{X\beta} + \varepsilon$, where y, ε are n vectors, $\boldsymbol{\beta}$ has dimension p, X has n×p, and $\varepsilon_i's$ are i.i.d. N(0,σ^2), show that the information matrix of $\boldsymbol{\beta}$ is $\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}$.

(20 Marks)

Question 5

If the goal for this software expert is to build an email spam filter: based on observed characteristics of an email message she wants to build a classification rule for assigning the message either as spam(marked with a "1") or not spam ("0"). To build the filter she has data of 4601 emails, and for each message she has a human-assigned to label 1 for spam, 0 for not spam, and the following characteristics:

- **caps_avg** = the average of the lengths of strings of capital letters used in the email (e.g. "The" = 1, "HELLO" = 5)
- **c_paren**, **c_exclaim**, **c_dollar** = the percentage of characters in the message which are parentheses ("(", "[", ")", "]"), exclamation point ("!"), and dollar sign ("$") respectively. (Percentages are between 0 and 100.)

a) In the current context, what are the two types of errors that a classifier can make? In the present context, is one type of mistake "worse" than the other? Explain your reasoning.

Use the following output to answer parts (b) - (e).

```
Call:
glm(formula = spam ~ caps_avg + c_paren + c_exclaim + c_dollar,
    family = "binomial", data = spam)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.75       0.07     -25   <2e-16 ***
caps_avg        0.21       0.02      12   <2e-16 ***
c_paren        -1.66       0.23      -7    2e-13 ***
c_exclaim       1.38       0.11      12   <2e-16 ***
c_dollar       11.86       0.62      19   <2e-16 ***

    Null deviance: 6170.2  on 4600  degrees of freedom
Residual deviance: 4160.7  on 4596  degrees of freedom
AIC: 4171

Number of Fisher Scoring iterations: 15
```

b) Provide a precise, numerical interpretation of the coefficient estimate for **c_dollar**. Do you and this result credible? Why or why not?

c) Provide a precise, numerical interpretation of the coefficient estimate for **caps_avg**. Do you find this result credible? Why or why not?

d) For a message that is 2% parentheses, 2% exclamation points, has zero dollar signs, and never strings together more than one capital letter, what estimated probability this message is spam?

e) If she uses the regression in above to build a classification rule based on the predicted probabilities. For some number K, we will flag a message as spam if the estimated P[spam = 1|X] > K. Referring to your answer in part (a), would you prefer to choose K = 1/4, K = 1/2, or K = 3/4? Why?

(4+3+3+5+5 Marks)

Question 6

a) If the hazard function is $h(t) = a\sqrt{t}$, where $a > 0$, what are the survival and density functions?

(10 Marks)

b) If survival times in the absence of censoring are distributed according to a Weibull distribution with parameters κ and λ, the hazard and survival functions can be written as

$$h(t) = \lambda\kappa t^{k-1}$$

$$S(t) = \exp(-\lambda t^{k})$$

respectively. If we observed data of the form (t_i, δ_i), where $\delta_i = 1$ if individual i fails at time t_i and $\delta_i = 0$ if i is right-censored at t_i, for i = 1,...,m. What is the log-likelihood function? Explan briefly how you might find the maximum-likelihood estimates of k and λ.

(10 Marks)

Question 7

150 female rats were bought in 50 litters of 3 and randomly given a placebo (2 rats per litter) or a new drug (1 rat per litter). The rats were followed for 4 months, and the time at which they developed tumours was recorded. Some rats died without developing tumours and were recorded as right-censored at the time of death. The following R commands and output have been used to test whether rats given the drug have the same survival function as those given the placebo.

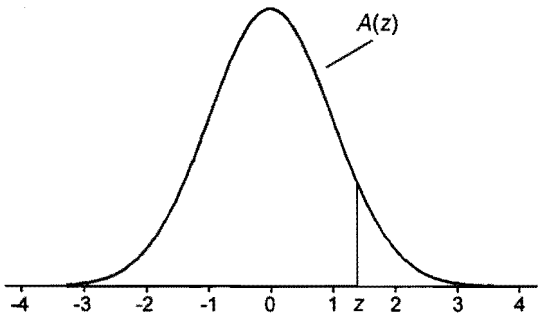
(20 Marks)

```
> survdiff(Surv(t,delta)~treat)
Call:
survdiff(formula = Surv(t, delta) ~ treat)

          N Observed Expected (O-E)^2/E (O-E)^2/V
treat=0 100       19     27.5      2.65       8.6
treat=1  50       21     12.5      5.86       8.6

 Chisq= 8.6  on 1 degrees of freedom, p= 0.00337


> coxph(Surv(t,delta)~treat)
Call:
coxph(formula = Surv(t, delta) ~ treat)


       coef exp(coef) se(coef)    z      p
treat 0.905      2.47    0.318 2.85 0.0044

Likelihood ratio test=7.97  on 1 df, p=0.00474  n= 150
```

What do you conclude?

TABLE A.1

Cumulative Standardized Normal Distribution



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

TABLE A.2

t Distribution: Critical Values of *t*

Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	Significance level 10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

TABLE A.4

χ^2 (Chi-Squared) Distribution: Critical Values of χ^2

	Significance level		
Degrees of freedom	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588