UNIVERSITY OF SWAZILAND



MAIN EXAMINATION PAPER 2017

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

In a prospective, randomized study to investigate the capacity of aspirin to prevent pregnancyinduced hypertension (=high blood pressure), 65 women were treated with a daily dose of either aspirin (34 women) or placebo (31 women) during the third trimester of pregnancy. The result is summarized in the Table 1 below;

	Hypertension	No hypertension	Total
Aspirin treatment	4	30	34
Placebo treatment	11	20	31
Total	15	50	65

Table 1: Contingency table for study comparing Aspirin and Placebo

In Table 2, the probability that the test statistic n_{11} from Fisher's exact test attains a specific value t is given (n_{11} = number of women treated with Aspirin with hypertension).

Table 2: Conditional probability that n_{11} attains value t for data in Table 1

t	0	1	2	3	4	5	6	7	
$P_{H_0}(n_{11}=t $	0	0	0.001	0.004	0.019	0.060	0.131	0.205	
t	. 8		9	10	11	12	13	14	15
$P_{H_0}(n_{11}=t $	0.230	0	.186	0.107	0.043	0.012	0.002	0	0

- a) Fisher's exact test evaluates the probability $P_{H_0}(n_{11} = t | ...)$ under a certain condition. What is this condition (what are the dots in this expression)? How are $P_{H_0}(n_{11} = t | ...)$ in Table 2 calculated? State a formula for this probability.
 - (2 p)
- b) Based on Table 2, calculate the one-sided p-value for Fisher's exact test for testing the null hypothesis that treatment and hypertension are independent versus the alternative that aspirin has a hypertension preventing effect. Interpret the result.

(3 p)

c) Consider a two-sided test for the independence null hypothesis versus the alternative that aspirin has any (positive or negative) effect on hypertension for pregnant women. There are different methods to define a two-sided p-value based on Fisher's exact test. A simple but uncommon way is to multiply the one-sided p-value by 2. Choose another method and describe exactly how the two-sided p-value is defined. Compute the two-sided p-value and interpret.

(3 p)

(

d) Considering again the one-sided test problem: Calculate the mid p-value and mention an advantage and a disadvantage of using mid p-values in contrast to usual p-values.

Consider the data in Question 1 but apply now large sample inference.

- a) Estimate the odds ratio θ and interpret it.
- b) Construct an approximate 95% CI for the odds ratio (Hint: an asymptotic expression for the variance of $log(\hat{\theta})$ is $\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{13}} + \frac{1}{n_{14}}$.
- c) Test the null hypothesis of independence versus the one-sided alternative that aspirin has a hypertension preventing effect.
- d) Use a test of your choice which uses a large sample distribution for the test statistic and report whether the two-sided test is significant at the 5% level.

(3 p)

(3 p)

(4 p)

(3 p)

Question 3

Let Y_1, \ldots, Y_n be independent random variables with binary distribution,

$$P(Y_i = 1) = p_i$$
 $P(Y_i = 0) = 1 - p_i$

a) We use a generalized linear model with identity link function g(x) = x and we assume that the probability p_i for the i-th observation to be 1 can be expressed as linear function depending on a explanatory variable x_i :

$$g(E(Y_i)) = E(Y_i) = \eta_i = \alpha + \beta x_i$$

(i) Determine the log likelihood function as function of the parameters α and β .

(5 marks)

(5

- (ii) Compute the likelihood equations for the ML-estimation of the parameters α and β . Mention how these equations are used to calculate the ML-estimates.
- b) The random variable y has a distribution in the exponential family, if its *p.d.f(or p.m.f)* can be written as

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

for a specific function $a(\phi)$, $b(\theta)$, and $c(y, \phi)$, where $a(\phi) > 0$ and $b(\theta)$ has up to twice derivatives. Let $l(\theta, \phi; y)$ denote associated log-likelihood function. It is known that $E\left(\frac{\delta l}{\delta \theta}\right) = 0$ and $Var\left(\frac{\delta l}{\delta \theta}\right) = -E\left(\frac{\delta^2 l}{\delta \theta^2}\right)$. Show that $E(Y) = b'(\theta)$ and $Var(Y) = a(\phi)b''(\theta)$. (10 Marks)

		Y	
X	Yes	No	- Total
Yes	n ₁₁	n ₁₂	n ₁₊
No	n ₂₁	$n_{22}^{}$	n_{2+}
Total	$n_{+1}^{}$	$n_{+2}^{}$	n

Suppose we have the following 2×2 table with all n_{ii} positive:

a) When X and Y are independent, given n, n_{1+}, n_{+1} , we have

$$p(n_{11}|n, n_{1+}, n_{+1}) = \frac{n_{1+}! n_{+1}! n_{2+}! n_{+2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!}$$

Show that

÷.,

$$E(n_{11}|n, n_{1+}, n_{+1}) = \frac{n_{1+} n_{+1}}{n}$$
(12 Marks)
b) Show that the sample odds ratio can be expressed as $\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$.
(8 Marks)

Question 5

An experiment was conducted to assess the efficacy of two species of insect predators (melanarius, tachyporus) in controlling aphids (an insect pest) on plants grown in separate insectproof cages. Observations were made over eight consecutive week-long periods, with fresh plants, aphid populations and predators being introduced each week. Eight plants were observed each week, with two plants receiving each of four treatment combinations (no predators, melanarius only, tachyporus only, both predators). In four of the eight weeks a fungicide was also applied to all the plants, to control outbreaks of mildew. It was expected that there might be differences in the development of aphid populations and the efficacy of the predators between the week-long periods because of variation in the temperature as well as the use of different batches of predators and aphids.

At the end of each week period, a count of the number of aphids was made for each plant. The output APPENDIX A shows the results of analysing the counts of aphids at the end of each week using a generalised linear model which allows overdispersion to be detected if it is present. The application of the fungicide to all the plants in particular weeks potentially imposes a more complex design structure, but a simplified structure was assumed to allow analysis using a generalised linear model. The fitted model allows for the main effects of each predator and the fungicide treatment, and for the interactions between these effects. The fitted model also includes terms accounting for differences between periods after allowing for fungicide effects, and differences in the separate and combined effects of the two predators between periods (again, after allowing for fungicide effects), these potentially contributing extra variability beyond the plant-to-plant variability against which treatment effects should be assessed.

a) Identify the components of a generalised linear model (GLM), showing how the expected value of the response is related to the explanatory model. Describe the particular form of GLM that is appropriate for these data.

; -

(5 Marks)

b) Briefly describe the method of iterative reweighted least squares used in fitting a generalised linear model.

(4 Marks)

c) Interpret the results of the analysis, in particular identifying the separate and combined efficacies of the two insect predators, and the impact of the fungicide application.

(8 Marks)

d) Describe what is meant by overdispersion, and indicate how evidence for overdispersion can be identified and tested for using the analysis output.

(3 Marks)

Question 6

a) Define the survivor function S(t) and the hazard function h(t) for a continuous random variable T measuring lifetime. Write down an expression for the survivor function in terms of the hazard function.

(3 Marks)

- b) The exponential distribution has constant hazard function $h(t) = \lambda$. Write down expressions for the density of the exponential distribution and the mean of this distribution in terms of λ . (2 Marks)
- c) Explain what is meant by a right-censored observation. Give two different examples of ways in which a right-censored observation might arise.

(3 Marks)

d) After a radical mastectomy for breast cancer, ten female patients were randomly assigned to one of two groups, an experimental group who received chemotherapy, and a control group who received no drugs. At the end of two years, survival times in months were recorded and are given in the table below. A right-censored observation is denoted by +, so 16+ denotes a right-censored observation at 16 months.

Experimental group	23	16+	18+	20+	24+
Control group	15	18	19	19	20

Compute the Kaplan-Meier estimate of the survivor function for each group and plot the results on one graph.

e) If survival times for the control group have an exponential distribution, estimate the hazard rate.

(2 Marks)

(10 Marks)

. . .

Consider the following experiment on visual perception using random-dot stereograms. A random-dot stereogram is composed of two rectangles placed side by side, where each rectangle appears to consist only of randomly scattered dots, without any image. When viewed with only one eye functioning, the viewer cannot see a hidden image. However, when viewed with both eyes, if a person focuses the eyes in front of or behind the pair of images, then a three-dimensional hidden image of a diamond can be seen. Sometimes the diamond image can be seen quickly, but on other occasions it can take a while before it can be perceived. Here the response variable is the time in seconds needed to perceive the diamond image. The experiment investigated whether giving a person prior knowledge about the shape of the image reduces the time needed to recognise it. Forty-three subjects (group NV) received just verbal information about the shape of the hidden object. Thirty-five subjects (group VV) received both verbal information and visual information, for example a drawing of the hidden object.

a) The response time T to perceive the diamond image has a Weibull distribution with hazard function

$$h(t) = \lambda \gamma t^{\gamma - 1}$$
 , $t \ge 0$

where $\lambda > 0$ and $\gamma > 0$ are parameters to be estimated. Show that the parameters are given by the intercept and slope of the theoretical relationship of the logarithm of the cumulative hazard function plotted against the logarithm of time.

(4 Marks)

b) Write down the proportional hazards model for a vector of p time-constant covariates, X, assuming a Weibull baseline hazard function. Interpret each term in your model. Write down an expression for the hazard ratio for this model.

(4 Marks)

c) Data from the random-dot stereogram experiment were analysed using a proportional hazards model with a Weibull baseline hazard function. One explanatory variable was included in the model: GroupVV, taking the value 1 if the subject was in Group VV, or 0 if the subject was in Group NV. The following edited computer output shows the results from the fitted model.

	Estimate	Standard Error
λ	0.060	0.019
γ	1.260	0.104
Group VV	0.552	0.233

(i) Use the fitted model to estimate the hazard ratio for a subject in the VV Group compared to a subject in the NV Group. Construct a 95% confidence interval for this hazard ratio. Which group, on average, has the shorter response times?

(4 Marks)

(ii) What is the estimated hazard function for a subject in Group NV? What is the estimated hazard function for a subject in Group VV? How does the estimated hazard function change with time for each group?

(4 Marks)

d) Someone suggested using an exponential distribution, instead of a Weibull distribution, to model the response times. What advice should you give regarding this idea? Justify your answer.

(4 Marks)

<u>APPENDIX</u> (Output for Question 5)

Accumulated analysis of deviance

Change	d.f.	deviance	mean deviance	deviance ratio	p-value
+ fungicide	1	23.91	23.91	1.88	0.179
+ tachyporus	1	990.58	990.58	78.08	<.001
+ melanarius	1	0.88	0.88	0.07	0.794
+ fungicide.period	6	1116.41	186.07	14.67	<.001
+ fungicide.tachyporus	1	183.47	183.47	14.46	<.001
+ fungicide.melanarius	1	4.63	4.63	0.36	0.550
+ tachyporus.melanarius	1	3.13	3.13	0.25	0.623
+ fungicide.tachyporus.period	6	516.48	86.08	6.79	<.001
+ fungicide.melanarius.period	6	108.37	18.06	1.42	0.236
+ fungicide.tachyporus.melanarius	1	0.01	0.01	0.00	0.975
+ fungicide.tachyporus.melanarius.period	6	197.20	32.87	2.59	0.037
Residual	32	405.97	12.69		
Total	63	3551.05	56.37		

Dispersion parameter is estimated to be 12.7 from the residual deviance The parameters estimate the effect of applying the fungicide or each predator

Estimates of some parameters

.

Parameter	estimate	s.e.	t value	<i>p</i> -value	antilog of estimate
Constant	5.218	0.185	28.14	<.001	184.5
fungicide yes	-0.455	0.298	-1.53	0.136	0.6341
tachyporus yes	-1.140	0.377	-3.03	0.005	0.3198
melanarius yes	0.108	0.255	0.42	0.676	1.114
fungicide yes.tachyporus yes	0.092	0.592	0.15	0.878	1.096
fungicide yes.melanarius yes	-0.099	0.416	0.24	0.813	0.9055
tachyporus yes.melanarius yes	0.425	0.486	0.87	0.388	1.529
fungicide yes.tachyporus yes.melanarius yes	0.458	0.750	0.61	0.545	1.581

Predicted means (standard errors)

.

;

			melanarius					
		٦	No	Y	es			
toohunoruo	No	128.34	(14.15)	131.19	(14.57)			
tachyporus	Yes	43.81	(8.18)	69.63	(11.27)			

		tachyporus				
	Γ	Ν	No	Yes		
funciaida	No	97.00	(11.11)	13.03	(2.85)	
lungicide	Yes	80.66	(9.72)	42.66	(6.96)	

		melanarius					
		N	0	Ye	es		
fungicide	No	54.78	(8.16)	55.25	(8.06)		
	Yes	72.62	(9.71)	50.69	(6.97)		

Table C-1. Cumulative Probabilities of the Standard Normal Distribution.

z(A) .09 x .00 .01 .02 .03 .04 .05 ,06 .07 .08 ,0 .5000 .5040 .5080 .5120 .5160 .5199 .5239 .5279 .5319 .5359 .5398 .5438 .5478 .5517 .5557 .5596 .5636 5675 .5714 .5753 .1 .2 .5793 .5832 .5871 .5910 ,\$948 .5987 .6026 .6064 .6103 .6141 .3 .6179 ,6217 .6255 .6293 .6331 .6368 .6406 6443 .6480 .6517 .6628 .6772 .6808 .6844 .4 .6554 .6591 ,6736 .6879 .6664 .6700 .6985 ,7088 .5 .6915 .6950 .7019 .7054 .7123 .7157 .7190 .7224 .6 .7257 .7291 .7324 .7357 .7389 .7422 .7454 .7486 .7517 7549 .7580 .7642 .7794 .7852 .7611 .7673 .7704 .7734 .7764 .7823 .8 7939 .7881 .7910 .7995 .2023 8051 .8078 .8106 .8133 .7967 .<u>.</u> .8340 ,8159 .8186 .8212 .8238 .8264 .8289 .8315 .8365 .8389 1.0 .8485 .8708 .8599 .8413 8438 .8508 .8531 .8577 .8621 .8461 .8554 .8790 .8665 .8749 .8944 .8810 8686 .8770 .8830 1.1 .8643 .8729 .8849 8869 8980 8997 1.2 8888 .8907 8925 .8962 .9015 1.3 .9082 .9177 .9032 .9049 .9066 .9115 .9131 .9147 .9099 .9162 .9207 .9292 .9306 1.4 .9192 .9222 .9236 .9251 .9265 .9279 .9319 .9332 .9345 .9357 .9406 .9418 .9429 .9441 1.5 .9370 .9382 .9394 ,9525 .9545 1,6 .9452 .9463 .9474 ,9484 .9495 .9505 .9515 .9535 .9573 ,9554 .9564 9582 .9591 1.7 .9599 .9608 .9616 .9625 .9633 1.8 .9641 .9649 .9656 .9664 .9671 .9678 .9686 .9693 .9699 .9706 1.9 .9713 .9719 .9726 .9732 .9738 .97.44 .9750 .9756 .9761 .9767 2.0 .9772 .9778 .9783 .9788 .9793 .9798 ,9803 .9808 .9812 .9817 2.1 .9821 .9826 ,9830 .9834 .9838 .9842 .9846 .9850 .9854 ,9857 2.2 .9861 9864 .9868 .987t .9875 .9878 .9881 .9884 .9887 .9890 2.3 .9893 9896 9898 .9901 .9904 .9906 .9909 .9911 9913 ,9916 .9918 .9920 .9922 ,9925 .9927 .9929 .9931 ,9932 .9934 .9936 2.5 .9938 .9940 .9941 .9943 .9945 .9946 .9948 .9949 .9951 .9952 2.6 .9953 9955 .9956 .9957 .9959 .9960 .9961 .9962 .9963 .9964 .9966 2.7 .9965 .9967 .9968 .9969 .9970 .9971 .9972 .9973 ,9974 .9977 2.8 ,9974 9975 .9976 .9977 .9978 9979 .9979 .9980 .9981 2.9 .9981 ,9982 .9982 .9983 .9985 .9986 .9986 .9984 .9984 .9985 3.0 .9987 ,9987 .9987 .9989 .9989 .9990 .9988 .9988 ,9989 .9990 3.1 3.2 .9990 .9991 .9991 .9991 .9992 9992 .9992 .9992 .9993 .9993 .9993 .9993 9994 .9994 9994 .9995 9995 .9995 .9994 9994 .9995 .9995 9995 .9996 .9996 3.3 .9996 .9996 .9996 .9996 ,9997 3.4 .9997 .9997 .9997 .9997 .9997 .9997 .9997 .9997 .9997 9998

Entry is area A under the standard normal curve from $-\infty$ to z(A)

Â

Volume II, Appendix C: page 2

Table C-2. Percentiles of the χ^2 Distribution



Volume II, Appendix C: page 3

.

Student's Distribution (t Distribution)

Table C-4 Percentiles of the t Distribution

Entry is $t(A; \nu)$ where $P\{t(\nu) \le t(A; \nu)\} = A$



		A				·		
4	.60	.70	.80	.85	.90	.95	.975	
<u> </u>	0.325	0.727	1.376	1.963	3.078	6.314	12.706	
2	0.289	0.617	1.061	1.386	1.886	2.920	4,303	
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	
6	0.265	0.553	0.906	L.134	1.440	1.943	2.447	
7	0.263	0.549	0.896	1.129	1.415	1.895	2.365	
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	
9	0.261	0.543	0.883	1.100	1.383	1.833	2,262	
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	
12	0.259	0.539	0.873	1.063	1.356	1.782	2.179	
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160	
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	
16	0.258	0.535	0.865	1.071	1.337	1.746	2,120	
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	
18	0.257	0.534	0.862	1.067	1,330	1.734	2,101	
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	
25	0.256	0.531	0.856	1.058	1.316	1,708	2.060	
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980	
00	0.253	0.524	0.842	1.036	1.282	1.645	1.960	

Volume II, Appendix C: page 6

	· A ·								
ν	.98	.985	.99	.9925	.995	.9975	.9995		
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590		
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598		
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924		
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610		
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869		
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959		
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408		
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041		
9	2.398	2.574	2.821	2,998	3.250	3.690	4.781		
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587		
1 1	2.328	2.491	2.718	2.879	3.106	3.497	4,437		
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318		
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221		
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140		
15	2.249	2.397	2.602	2.745	2.947	3.286	4.073		
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015		
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965		
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922		
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883		
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849		
21	2.189	2.328	2.518	2.649	2,831	3.135	3.B19		
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792		
23	2.177	2.313	2.500	2.629	2.807	3,104	3,768		
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745		
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725		
26	2.162	2.296	2,479	2.605	2.779	3.067	3.707		
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690		
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674		
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659		
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646		
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551		
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460		
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373		
00	2.054	2.170	2.326	2.432	2.576 ·	2.807	3.291		

Table C-4 (Continued) Percentiles of the t Distribution

Volume II, Appendix C: page 7

.