JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF HEALTH SCIENCES

UNIVERSITY EXAMINATION FOR DEGREE OF MASTER PUBLIC HEALTH

1ST YEAR 2ND SEMESTER 202023/2024 ACADEMIC YEAR

KISUMU LEARNING CENTRE

COURSE CODE:	**HMP 5123**
COURSE TITLE:	**ADVANCED BIOSTATISTICS**
EXAM VENUE:	**STREAM:**
DATE:	**EXAM SESSION:**
TIME:	**3.00 HOURS**

Instructions:

1. **Answer any four Questions (Question One is Compulsary)**
2. **Candidates are advised not to write on the question paper.**
3. **Candidates must hand in their answer booklets to the invigilator while in the examination room.**

SECTION A

Answer question one(Compulsary)

1. Question one (10 marks).

a) Show that $se(\hat{\pi}) = \frac{1}{\sqrt{-l''(\hat{\pi})}}$ where $se(\hat{\pi})$ is the standard error of the estimated population prevalence and $l''(\hat{\pi})$ is the second derivative of the log-likelihood function of the estimated population prevalence. (5 marks)

b) Derive the wool's formulae for log odds $se[\log(\hat{\Omega})] = \sqrt{\frac{1}{x} + \frac{1}{n-x}}$ for X being the number of individuals with disease, $se[\log(\hat{\Omega})]$ is the standard error of the log-odds and n is the sample size. (5 marks)

SECTION B

Answer any three Questions

2. Question five (20 marks).

The table below gives results of 6 clinical trials comparing the risk of OHSS (ovarian hyperstimulation syndrome) between recombinant FSH and urinary FSH used during an IVF (in vitro fertilization) treatment.

Trial	No. of patients Rec FSH	No. of patients Ur FSH	OHSS Rec FSH	OHSS Ur FSH
1	585	396	19	8
2	57	33	3	0
3	54	35	2	1
4	119	114	6	2
5	60	63	2	1
6	105	67	8	3

A meta-analysis was carried out using Mantel-Haenszel's procedure, stratified on trial. Some SPSS output is given at the following pages. Read this output and answer the following questions.

Risk Estimate

Trial		Value	95% Confidence Interval Lower	95% Confidence Interval Upper
1	Odds Ratio for FSH (Recombinant / Urinary)	.614	.266	1.417
	For cohort OHSS = no	.987	.967	1.008
	For cohort OHSS = yes	1.608	.711	3.636
	N of Valid Cases	981		
2	For cohort OHSS = no	.947	.891	1.007
	N of Valid Cases	90		
3	Odds Ratio for FSH (Recombinant / Urinary)	.765	.067	8.765
	For cohort OHSS = no	.991	.918	1.071
	For cohort OHSS = yes	1.296	.122	13.763
	N of Valid Cases	89		
4	Odds Ratio for FSH (Recombinant / Urinary)	.336	.066	1.702
	For cohort OHSS = no	.967	.921	1.014
	For cohort OHSS = yes	2.874	.592	13.947
	N of Valid Cases	233		
5	Odds Ratio for FSH (Recombinant / Urinary)	.468	.041	5.297
	For cohort OHSS = no	.982	.928	1.039
	For cohort OHSS = yes	2.100	.195	22.561
	N of Valid Cases	123		
6	Odds Ratio for FSH (Recombinant / Urinary)	.568	.145	2.223
	For cohort OHSS = no	.967	.897	1.043
	For cohort OHSS = yes	1.702	.468	6.188
	N of Valid Cases	172		

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)

Breslow-Day	1.507	5	.912
Tarone's	1.507	5	.912

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.513
ln(Estimate)			-.668
Std. Error of ln(Estimate)			.308
Asymp. Sig. (2-sided)			???
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	???
		Upper Bound	???
	ln(Common Odds Ratio)	Lower Bound	???
		Upper Bound	???

.

a. Make a 2X2 table for the first trial.
    i. Compute the OHSS odds ratio of recombinant FSH treatment relative to urinary FSH treatment. (1 mark)
    ii. Compute also the corresponding relative risk (1 mark)
    iii. How are these estimates related to the estimates given for trial 1 in the first table of the SPSS output (1 mark)
    iv. What is the difference between the two relative risk estimates? (1 mark)
b. Give the OHSS odds ratios of recombinant FSH relative to urinary FSH per trial. Is the assumption that the true odds ratios are equal across trials warranted? (7 marks)
    i. Motivate your answer. (1 mark)
c. Give the Mantel-Haenzel estimate of the common OHSS odds ratios of recombinant FSH relative to urinary FSH. (2 marks)
    i. Is it justified to interpret it as a relative risk? (1 mark)
d. Fill in the question marks in the third table. (5 marks)

3. Question two (20 marks).

In a randomized trial patients infected by helicobacter pylori were randomly allocated to treatment by drug combination A or treatment by drug combination B. At the end of the study, the non-cure rates are to be compared between the two groups, using the risk difference or the risk ratio as effect measure.

RESIST resistant against one of the drugs in the combination

0 = no, 1 = yes

CURE 1 = not cured, 0 = cured

TREAT 0 = drug combination A , 1 = drug combination B

treat * cure * resist Crosstabulation

Count

resist			cure: cured	cure: not cured	Total
no	treat	A	111	3	114
		B	99	6	105
	Total		210	9	219
yes	treat	A	90	9	99
		B	75	12	87
	Total		165	21	186
Total	treat	A	201	12	213
		B	174	18	192
	Total		375	30	405

Some SPSS output

treat * cure Crosstabulation

			cure: cured	cure: not cured	Total
treat	A	Count	201	12	213
		% within treat	94.4%	5.6%	100.0%
	B	Count	174	18	192
		% within treat	90.6%	9.4%	100.0%
Total		Count	375	30	405
		% within treat	92.6%	7.4%	100.0%

Risk Estimate

	Value	95% Confidence Interval: Lower	95% Confidence Interval: Upper
Odds Ratio for treat (A / B)	1.733	.812	3.698

For cohort cure = cured	1.041	.984	1.101
For cohort cure = not cured	.601	.297	1.215
N of Valid Cases	405		

Generalized Linear Models

Model Information

Dependent Variable	cure[a]
Probability Distribution	Binomial
Link Function	Identity

a. The procedure models not cured as the response, treating cured as the reference category.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.056	.0158	.025	.087	12.716	1	.000
[treat=1.00]	.037	.0263	-.014	.089	2.022	1	.155
[treat=.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: cure

Model: (Intercept), treat

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Generalized Linear Models

Model Information

Dependent Variable	cure[a]
Probability Distribution	Binomial
Link Function	Log

a. The procedure models not cured as the response, treating cured as the reference category.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	Hypothesis Test df	Hypothesis Test Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	-2.876	.2804	-3.426	-2.327	105.210	1	.000	.056	.033	.098
[treat=1.00]	.509	.3591	-.195	1.213	2.011	1	.156	1.664	.823	3.364
[treat=.00]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1[b]									

Dependent Variable: cure

Model: (Intercept), treat

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

treat * cure * resist Crosstabulation

resist				cure: cured	cure: not cured	Total
no	treat	A	Count	111	3	114
			% within treat	97.4%	2.6%	100.0%
		B	Count	99	6	105
			% within treat	94.3%	5.7%	100.0%
	Total		Count	210	9	219
			% within treat	95.9%	4.1%	100.0%
yes	treat	A	Count	90	9	99
			% within treat	90.9%	9.1%	100.0%
		B	Count	75	12	87
			% within treat	86.2%	13.8%	100.0%
	Total		Count	165	21	186
			% within treat	88.7%	11.3%	100.0%
Total	treat	A	Count	201	12	213

		% within treat	94.4%	5.6%	100.0%
	B	Count	174	18	192
		% within treat	90.6%	9.4%	100.0%
Total		Count	375	30	405
		% within treat	92.6%	7.4%	100.0%

Risk Estimate

resist		Value	95% Confidence Interval Lower	Upper
no	Odds Ratio for treat (A / B)	2.242	.546	9.205
	For cohort cure = cured	1.033	.977	1.092
	For cohort cure = not cured	.461	.118	1.795
	N of Valid Cases	219		
yes	Odds Ratio for treat (A / B)	1.600	.640	4.003
	For cohort cure = cured	1.055	.950	1.171
	For cohort cure = not cured	.659	.292	1.489
	N of Valid Cases	186		
Total	Odds Ratio for treat (A / B)	1.733	.812	3.698
	For cohort cure = cured	1.041	.984	1.101
	For cohort cure = not cured	.601	.297	1.215
	N of Valid Cases	405		

Answer the following questions

a. Compare the risk of not getting cured with either treatment A or B. Find the risk difference/risk ratio and their confidence intervals for the non-cure rate. (hint transformation cure=1). (4 marks)
b. Compute the Wald's 95%CI. (4 marks)
c. What is the interpretation of the estimated regression coefficients from the first and second regression model? (2 marks)
d. Some of the patients are resistant to one of the drugs in the drug combination, others are not resistant. Adjust the estimates of the risk difference for resistance. (2 marks)
e. Compute the risk differences and risk ratio in both strata. (4 marks)
f. Calculate the weighted mean of the two risk difference using the weight factor for each stratum: *One over the squared standard error*. (4 marks)

4. Question three (20 marks).

Presence of a certain element of the set of teeth in babies, depending on age: Y=1/0 if element present/absent, X=age at examination (weeks). Using binary logistic regression in SPSS gives the following:

Dependent Variable Encoding

Original Value	Internal Value
no	0
yes	1

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	55.226	-1.040
	2	55.108	-1.150
	3	55.108	-1.153
	4	55.108	-1.153

a. Constant is included in the model.
b. Initial -2 Log Likelihood: 55.108
c. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a,b]

	Observed		Predicted element no	yes	Percentage Correct
Step 0	element	no	38	0	100.0
		yes	12	0	.0
	Overall Percentage				76.0

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.153	.331	12.117	1	.000	.316

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Age in weeks	20.805	1	.000
	Overall Statistics		20.805	1	.000

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	Age in weeks
Step 1	1	36.215	-3.827	.095
	2	29.677	-6.483	.162
	3	27.743	-8.796	.220

	4	27.474	-10.043	.251
	5	27.467	-10.287	.257
	6	27.467	-10.295	.257
	7	27.467	-10.295	.257

a. Method: Enter
b. Constant is included in the model.
c. Initial -2 Log Likelihood: 55.108
d. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27.467[a]	.425	.636

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

	Observed		Predicted element no	yes	Percentage Correct
Step 1	element	no	34	4	89.5
		yes	3	9	75.0
	Overall Percentage				86.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Age in weeks	.257	.078	10.727	1	.001	1.293	1.109	1.508
	Constant	-10.295	3.066	11.275	1	.001	.000		

a. Variable(s) entered on step 1: Age in weeks.

Correlation Matrix

		Constant	Age in weeks
Step 1	Constant	1.000	-.987
	Age in weeks	-.987	1.000

a. Estimate of the covariance matrix, hence what are the standard errors (s_0) and (s_1)? (4 marks)
b. What is the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$. (2 marks)
c. Give the 95%CI for β_1 using the Wald's method. (2 marks)
d. What is the precision of the estimated probability $\hat{\pi}(x)$? (4 marks)
e. What is the probability that a 40 week old will have the element? Give 95%CI for this probability. (5 marks)
f. Test for $H_0: \beta_1 = 0$ with three methods (follow SPSS output). (3 marks)

5. Question four (20 marks).

In a random sample from the population of a rural area in a certain developing country the following variables, among others, were observed on 328 persons. SYS=systolic blood pressure (mmHg), PULSE=pulse rate (beats/min), and SES=social economic status (1=lower class, 2=middle class, 3=upper class)

This problem concentrates on the differences in mean systolic blood pressure between the three social economic classes corrected for pulse frequency. Three multiple regression models were filled using SPSS. Part of the output is given below.

Model 1:

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	middle social economic class, low social economic status(a)	.	Enter

a All requested variables entered.
b Dependent Variable: systolic blood pressure (mmHg)

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4019.437	2	2009.719	6.898	.001(a)
	Residual	94683.840	325	291.335		
	Total	98703.277	327			

a Predictors: (Constant), middle social economic class, low social economic status
b Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	126.381	1.002		126.175	.000
	low social economic status	-2.196	1.307	-.095	-1.681	.094
	middle social economic class	-3.645	1.330	-.155	-2.741	.006

a Dependent Variable: systolic blood pressure (mmHg)

Model 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.169(a)	.029	.026	17.15045
2	.258(b)	.067	.058	16.86281
3	.259(c)	.067	.055	16.88669

a Predictors: (Constant), pulse frequency (beats/min)
b Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class
c Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class, squared pulse rate

ANOVA(d)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2814.288	1	2814.288	9.568	.002(a)
	Residual	95888.989	326	294.138		
	Total	98703.277	327			
2	Regression	6572.496	3	2190.832	7.705	.000(b)
	Residual	92130.781	324	284.354		
	Total	98703.277	327			
3	Regression	6596.497	4	1649.124	5.783	.000(c)
	Residual	92106.780	323	285.160		
	Total	98703.277	327			

a Predictors: (Constant), pulse frequency (beats/min)
b Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class
c Predictors: (Constant), pulse frequency (beats/min), low social economic status, middle social economic class, squared pulse rate
d Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	104.616	6.711		15.588	.000
	pulse frequency (beats/min)	.250	.081	.169	3.093	.002
2	(Constant)	106.752	6.625		16.113	.000
	pulse frequency (beats/min)	.239	.080	.161	2.996	.003
	low social economic status	-2.196	1.291	-.095	-1.701	.090
	middle social economic class	-3.472	1.315	-.147	-2.641	.009
3	(Constant)	97.588	32.277		3.024	.003
	pulse frequency (beats/min)	.464	.781	.313	.594	.553
	low social economic status	-2.223	1.296	-.096	-1.715	.087
	middle social economic class	-3.432	1.324	-.146	-2.592	.010
	squared pulse rate	-.001	.005	-.153	-.290	.772

a Dependent Variable: systolic blood pressure (mmHg)

Model 3:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.258(a)	.067	.058	16.86281
2	.284(b)	.081	.066	16.78795

a Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status
b Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status, mid_pulse, low_pulse

ANOVA(c)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6572.496	3	2190.832	7.705	.000(a)
	Residual	92130.781	324	284.354		
	Total	98703.277	327			
2	Regression	7952.374	5	1590.475	5.643	.000(b)
	Residual	90750.903	322	281.835		
	Total	98703.277	327			

a Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status

b Predictors: (Constant), middle social economic class, pulse frequency (beats/min), low social economic status, mid_pulse, low_pulse
c Dependent Variable: systolic blood pressure (mmHg)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	106.752	6.625		16.113	.000
	pulse frequency (beats/min)	.239	.080	.161	2.996	.003
	low social economic status	-2.196	1.291	-.095	-1.701	.090
	middle social economic class	-3.472	1.315	-.147	-2.641	.009
2	(Constant)	101.067	7.155		14.125	.000
	pulse frequency (beats/min)	.306	.086	.207	3.568	.000
	low social economic status	12.841	9.568	.554	1.342	.181
	middle social economic class	6.729	9.051	.285	.743	.458
	low_pulse	-.181	.115	-.653	-1.578	.116
	mid_pulse	-.123	.109	-.435	-1.130	.259

a Dependent Variable: systolic blood pressure (mmHg)

In order to look at the crude differences in mean systolic blood pressure between the three groups, model 1 is fitted. Study the output of model 1, notice in particular how the independent variables are coded (***LOW:1 0 -1 and MID: 0 1 -1***), and answer questions (a) to (d).

e. What is the interpretation of the estimated regression coefficients of the independent variables "low social economic status" and "middle social economic status"? (1 mark)
   i. Give also the interpretation of the estimated intercept. (1 mark)
   ii. Compute the estimates for the mean systolic blood pressure of the three SES classes. (1 mark)
f. Are there significant differences in mean systolic blood pressures between the SES groups?
   i. Formulate the null hypothesis and give the p-value. (1 mark)
g. Give the estimate of the within groups standard deviation of systolic blood pressure. (1 mark)
   i. How can this be used to compute an (approximate) 95% confidence interval for the group means? (1 mark)

   ii. Give this confidence interval for the low SES group. (the number of individuals in the lower SES group was 138) (1 mark)

h. Give the estimate of the percentage variability in systolic blood pressure that is explained by differences between SES classes. (1 mark)

In order to look at the differences in mean systolic blood pressure between the SES groups corrected for pulse rate, model 2 was fitted. Study the output of model 2 and answer the questions (e) and (h).

i. Are there significant differences in mean systolic blood pressures between the SES groups corrected for pulse rate? (1 mark)
   i. Formulate the null hypothesis and give the p-value. (2 marks)
j. Give the estimate of the pulse rate corrected difference in mean systolic blood pressure between the low and middle SES group. (1 mark)
   i. Do the same for the low and high group and for the middle and high group. (1 mark)
k. Compute the estimate, based on model 2, of the mean systolic blood pressure for middle class people with pulse rate equal to 70 (1 mark)
l. One of the assumptions underlying model 2 is that the relation between systolic blood pressure and pulse rate is linear. Is this assumption reasonable in this case? (1 mark)
   i. Motivate your answer. (1 mark)

One of the assumptions of model 2 is that there is no interaction between SES classes and pulse rate. In order to investigate whether this assumption is justified, model 3 was fitted. Study the output of model 3 and answer the following questions.

m. Is there statistical evidence that there is interaction between SES class and pulse rate? (1 mark)
   i. Motivate your answer. (1 mark)
n. Give the equation of the estimated regression line (based on model 3) of systolic blood pressure against pulse rate for the low SES group. (1 mark)

What is the estimated difference (based on model 3) in mean systolic blood pressure (1 mark)