WI-2-60-1-6

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY
UNIVERSITY EXAMINATIONS
2024/2025 ACADEMIC YEAR

MASTER OF SCIENCE IN EPIDEMIOLOGY AND BIOSTATISTICS

PEH3105: BIOSTATISTICS FOR EPIDEMIOLOGICAL STUDIES

DATE: DECEMBER 2024 **TIME: 3 HOURS**

INSTRUCTIONS: ANSWER ANY FOUR QUESTIONS

1)

(a) In designing epidemiological studies:

(i) Explain why sample size calculation is important. *(4 Marks)*

(ii) What is the *p*-value, the significance level and the power of a test? *(6 Marks)*

(b) In Kenya, the distribution of body mass index (BMI) in people between 40 to 60 years is approximately normal with mean $\mu = 23.5 kg/m^2$ and standard deviation $\sigma = 7.5 kg/m^2$. A master's student plans to conduct a study at the KNH Medical Clinic to determine if patients in the same age-group newly diagnosed with hypertension have the same mean. If the true mean BMI in these patients is as low as $22 kg/m^2$ (or as high as $25 kg/m^2$) the student wants a risk of only 10% of failing to reject the null hypothesis. If the student planned to conduct a two-sided test at the 0.05 significant level what sample size is needed for this study? *(6 Marks)*

$$\mathbf{n} = \frac{\sigma^2 (Z_{\alpha/2} + Z_{1-\beta})^2}{d^2} \; ; \; (Z_{\alpha/2} = 1.96 \; and \; Z_{1-\beta} = 1.28)$$

(c) In a bladder cancer study, 210 randomly selected men attending a clinic in Nairobi were asked about their cigarette smoking habits (cigarette sticks per day). The results of the study indicated an average consumption of 13 cigarettes per day.

(i) Using a population standard deviation of 9 cigarettes per day, construct a 95% confidence interval for the cigarette smoking in this population. *(4 Marks)*

(ii) Would there be a significant difference if the same study had been done in 310 randomly selected men attending a similar clinic in Nakuru where the average consumption was 11 cigarettes per day? *(5 Marks)*

$$CI = \bar{x} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}; \bar{x} + Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

2)

(a) State the appropriate multivariable regression technique that would be used to analyse each of the following studies:

(i) A randomised-controlled trial of Typhoid Vi vaccine in 2- to 16-year-old children to prevent typhoid infections. *(2 Marks)*

(ii) A retrospective cohort study to determine if birthweight is associated with the birth order of the baby (1st-born, 2nd-born etc.) and the age of the mother. *(2 Marks)*

(iii) A case-control study of the association between five risk factors (age, sex, level of education, HIV infection and length of TB treatment) and the development of multidrug resistant TB. *(2 Marks)*

(iv) A cross-sectional study of the association between consumption of khat *(Catha edulis)* and depression in youth aged 18 to 35 years. *(2 Marks)*

(v) A prospective cohort study in women to determine if the circumcision status of their male partner is a risk factor for pelvic inflammatory disease (PID). *(2 Marks)*

(b) In a study to determine the effect of various factors on the development of subsequent thromboembolic events (including pulmonary thrombosis) in women newly diagnosed and on treatment for deep venous thrombosis (DVT), a Cox Proportional Hazards model below was constructed where the effects of history of contraceptive use (hx_oc, 1= history of use, 0=no history of use) was examined. Age (years), weight (kg), presence of varicose veins (0=no, 1=yes), physical activity (0=none 1=physically active) and alcohol consumption (0=no, 1=yes) were also examined in a multivariate model. [NB/ follow-up time was in days for 3-years]:

```
. stcox hx_oc age weight varicose p_activity alcohol

No. of subjects =          208                 Number of obs    =         208
No. of failures =           28
Time at risk    =       136270
                                                LR chi2(6)       =       56.60
Log likelihood  =   -86.316123                  Prob > chi2      =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hx_oc	4.80515	3.30929	2.28	0.023	1.245912	18.53218
age	1.10888	.0486788	2.35	0.019	1.017461	1.208514
weight	.9655006	.0233608	-1.45	0.147	.920783	1.01239
varicose	4.148015	3.075786	1.92	0.055	.969767	17.74243
p_activity	.6365411	.275617	-1.04	0.297	.2724352	1.48727
alcohol	4.805787	2.201294	3.43	0.001	1.958263	11.79391

(i) What was the effect of history of oral contraceptive use on the development of subsequent thromboembolic events after diagnosis with DVT? *(3 Marks)*

(ii) What was the effect of varicose veins on the development of subsequent thromboembolic events after diagnosis with DVT? *(3 Marks)*

(iii) Is there evidence of confounding by the age and weight of the women enrolled in this study? *(3 Marks)*

(iv) Was there overfitting in this model? Why or why not? *(3 Marks)*

(v) Outline two other confounding variables which you might consider in this study? *(3 Marks)*

3) The Western Collaborative Group Study (WCGS) was a prospective cohort study of 3,154 initially well men, aged 39-59 years at enrolment in 1960-61, who were employed in ten participating companies in California, USA. The effect of total cholesterol (mg/dl), age (years), weight (kg), cigarette smoking (no. per day) on systolic blood pressure (sbp) was assessed. The following is a STATA output of the simple linear regression model on the association between total cholesterol (predictor) and sbp (outcome) in these men:

- `regress sbp cholesterol`

Source	SS	df	MS		
Model	10778.0142	1	10778.0142	Number of obs	= 3,142
Residual	700920.042	3,140	223.222943	F(1, 3140)	= 48.28
				Prob > F	= 0.0000
Total	711698.056	3,141	226.583272	R-squared	= 0.0151
				Adj R-squared	= 0.0148
				Root MSE	= 14.941

sbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cholesterol	.042662	.0061396	6.95	0.000	.0306239	.0547001
_cons	118.9517	1.415168	84.05	0.000	116.1769	121.7264

(a) Briefly describe the relationship between total cholesterol and systolic blood pressure (sbp) using the results in this model. *(3 Marks)*

(b) Explain the meaning of the following outputs in the model:

(i) Cholesterol (P>| t | = 0.000) *(2 Marks)*

(ii) Coefficient (_cons = 118.9517) *(2 Marks)*

(iii) Probability (Prob > F = 0.0000) *(2 Marks)*

(c) Which two ways could be used to evaluate the fit of this model? *(4 Marks)*

(d) What assumptions were made in using the method of least squares to estimate the population line in this model? *(4 Marks)*

The following is a STATA output of the multiple regression model for the association between sbp (outcome) and total cholesterol and several other predictor variables:

- `regress sbp cholesterol age weight cigarettes`

Source	SS	df	MS		
Model	76740.7624	4	19185.1906	Number of obs	= 3,142
Residual	634957.294	3,137	202.409083	F(4, 3137)	= 94.78
				Prob > F	= 0.0000
Total	711698.056	3,141	226.583272	R-squared	= 0.1078
				Adj R-squared	= 0.1067
				Root MSE	= 14.227

sbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cholesterol	.0352444	.0058987	5.97	0.000	.0236787	.0468101
age	.4632338	.0462186	10.02	0.000	.3726121	.5538556
weight	.1853026	.0120929	15.32	0.000	.1615918	.2090133
cigarettes	.0436406	.0176638	2.47	0.014	.0090069	.0782743
_cons	67.19822	3.225457	20.83	0.000	60.874	73.52244

(e) Briefly describe the relationship between cigarette smoking and sbp using the results in this multiple regression model? *(4 Marks)*

(f) Why do you think the coefficient of blood glucose reduced from 0.042 in the simple regression model to 0.035 in the multiple regression model? *(4 Marks)*

4) A study was done in infants to assess the risk of sudden infant death syndrome (SIDS) [which is the sudden death of a seemingly healthy infant under 1 year of age that cannot be explained following a thorough case investigation]. The usual sleeping position (0=baby placed to sleep on his back, 1=baby placed to sleep on his side or stomach), the gestational age of the baby at birth (in weeks), the birth weight (in grams), mother's age at delivery (in years) and mother's cigarette smoking habits during pregnancy (0=no smoking, 1=smoker) were examined. The following STATA output represents the results of this study showing the association between these predictors and sudden infant death syndrome (SIDS) in these infants:

```
. logit sids_death sleeping_p gestage birthwt momage cigarette_pg

Iteration 0:   log likelihood = -75.060364
Iteration 1:   log likelihood = -60.778044
Iteration 2:   log likelihood = -58.325289
Iteration 3:   log likelihood = -58.297612
Iteration 4:   log likelihood = -58.297601
Iteration 5:   log likelihood = -58.297601

Logistic regression                             Number of obs     =        150
                                                LR chi2(5)        =      33.53
                                                Prob > chi2       =     0.0000
Log likelihood = -58.297601                     Pseudo R2         =     0.2233
```

sids_death	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sleeping_p	1.153952	.5820275	1.98	0.047	.0131986	2.294704
gestage	.5102449	.1318062	3.87	0.000	.2519096	.7685803
birthwt	-.0003793	.0006495	-0.58	0.559	-.0016523	.0008937
momage	-.1948352	.0630103	-3.09	0.002	-.3183331	-.0713374
cigarette_pg	.6200056	.5280774	1.17	0.240	-.415007	1.655018
_cons	-14.2035	4.404296	-3.22	0.001	-22.83576	-5.571239

(a) What is the relationship between the sleeping position of the infant and SIDS? *(5 Marks)*

(b) Interpret the z statistic and associated *P-value* for the sleeping position. *(4 Marks)*

(c) Interpret the likelihood ratio chi-squared test (LR chi^2) provided for the model. *(4 Marks)*

(d) If a particular infant was born weighing 2500 grams at a gestational age of 32 weeks to a 20-year-old mother who smoked during pregnancy and but had learnt to place the baby on her back when putting her to sleep, what is the probability that the baby would experience SIDS? *(8 Marks)*

(e) What is the odds ratio of developing SIDS for infants born of mothers who smoked during pregnancy? *(4 Marks)*

5) A PhD student conducted a study to evaluate the effect of parasitic infection on the white blood cell differential count (a test which measures the percentage of each type of white blood cells in blood – neutrophils, lymphocytes, monocytes, basophils, eosinophils etc.). Five hundred and thirty (530) children between 0 to 10 years were enrolled in the study and their blood taken for WBC differential count and a stool sample was taken and analysed for parasitic infection. Background characteristics (e.g., age, sex, SES) were also taken. A two-sided t-test was carried out to determine the difference in mean monocyte count between those diagnosed to have parasitic infection (positive) and those who were negative (normal monocyte count ranges from 2% to 8% of total white blood cell count). The results are shown in the table below:

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
negative	335	7.389761	.181424	3.320604	7.032884	7.746639
positive	195	6.503077	.1989858	2.778685	6.110624	6.89553
combined	530	7.063528	.1371997	3.158574	6.794005	7.333051
diff		.8866843	.2821474		.332415	1.440953

diff = mean(negative) - mean(positive) t = 3.1426
Ho: diff = 0 degrees of freedom = 528

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9991 Pr(|T| > |t|) = 0.0018 Pr(T > t) = 0.0009

(a) Interpret these t-test results. *(8 Marks)*

Linear regression analysis model was created to find out if the effect of parasitic infection (positive=1, negative=0) on the percentage of monocytes in blood was influenced by the age (in months) and the sex of the child (sex: 0=female, 1=male). The model is given here below:

Source	SS	df	MS		
Model	229.124069	3	76.3746897	Number of obs	530
Residual	5048.49247	526	9.59789443	F(3, 526)	7.96
Total	5277.61654	529	9.97659081	Prob > F	0.0000
				R-squared	0.0434
				Adj R-squared	0.0380
				Root MSE	3.098

monocytes	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
parasite	-.7176868	.2857257	-2.51	0.012	-1.278991	-.1563832
agemonths	-.0126054	.0045447	-2.77	0.006	-.0215334	-.0036773
sex	-.7076012	.2696117	-2.62	0.009	-1.237249	-.1779533
_cons	8.483448	.3491876	24.29	0.000	7.797474	9.169422

(b) Interpret the results in this model as to the effect of age in months and sex of the children under study on mean percentage of monocytes in blood (table above). *(7 Marks)*

(c) The significance test in the linear regression above for the association between percentage monocytes and the age in months uses the *t-distribution with n-2 degrees of freedom*. Why is this and what are the two restrictions? *(4 Marks)*

A t-test was then carried out to further assess the relationship between age in months and parasitic infection (negative or positive) (see table below):

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
negative	335	58.44478	1.636731	29.95709	55.22518	61.66438
positive	195	71.95897	2.098183	29.29953	67.8208	76.09715
combined	530	63.41698	1.320373	30.39727	60.82316	66.0108
diff		-13.5142	2.676737		-18.77256	-8.255836

diff = mean(negative) - mean(positive) t = -5.0488

Ho: diff = 0 degrees of freedom = 528

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

(d) Comment on the results of this t-test in relation to the previous two tables (above). *(6 Marks)*

6) While your colleague is evaluating the usage of various "statins" (drugs that are conventionally used to treat high cholesterol) in a large administrative database sent to you by your epidemiology lecturer, she finds what she thinks is an association between the statins and breast cancer. She does one more analysis asking whether there is perhaps interaction by pre- versus post-menopausal status. She stratifies the data using pre- and post-menopausal status and ends up with three tables as follows:

Crude Data

All Women	Breast Cancer	
	Yes	No
Statins	19	1393
No statins	29	1376

Stratum Specific Data

Pre-Menopausal Women

	Breast Cancer	
	Yes	No
Statins	8	205
No Statins	5	200

Post-Menopausal Women

	Breast Cancer	
	Yes	No
Statins	11	1188
No Statins	24	1176

(a) Calculate the crude risk ratio and the two stratum-specific risk ratios *(show your work). (9 Marks)*

(b) Calculate the Mantel-Haenszel adjusted risk ratio *(use the formula below and show your work). (5 Marks)*

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

She then uses STATA to generate a Mantel-Haenszel adjusted risk ratio model below:

```
. cs breast_cancer statins, by( menopausal)
```

menopausal	RR	[95% Conf.	Interval]	M-H Weight
Pre-Menopausal	1.539906	.5121943	4.629709	2.547847
Post-Menopausal	.4587156	.2257189	.9322212	11.995
Crude	.6519244	.3672991	1.15711	
M-H combined	.6481357	.364752	1.151687	

Test of homogeneity (M-H) chi2(1) = 3.287 Pr>chi2 = 0.0698

(c) Would you conclude that the menopausal status is a confounder in the relationship between statins use and breast cancer? Why or why not? *(4 Marks)*

(d) Would you conclude that the menopausal status is an effect modifier in the relationship between statins and breast cancer? Why or why not? *(4 Marks)*

(e) State in clear terms how menopause status affects the relationship between statins use and breast cancer in these women. *(3 Marks)*