MSc II: **Medical Statistics** (Spring Term 11/12)

Assignment Submission Date: **Wednesday 02 May 2012**

1. (a) Consider a stratified contingency table analysis to measure the association between a disease (or other outcome) and an exposure, where the i-th (2×2) table of K is given by

	Disease status		
	Yes	No	Total
Exposed	a_i	b_i	n_{1i}
not Exposed	c_i	d_i	n_{0_i}
			N_i

where $N_i = n_{0i} + n_{1i}$ is the total sample size for each table.

Recall that the estimated relative risk from each table is then

$$\widehat{RR}_i = \hat{p}_1/\hat{p}_0 = \frac{a_i/n_{1i}}{c_i/n_{0i}}$$

(i) Describe briefly what the relative risk represents in probabilistic terms. Why is it usual when constructing a confidence interval for the relative risk to use the standard error of $\log(\widehat{RR})$ rather than of $\widehat{RR}$ directly? Write down an estimate of $\sigma_i^2 = \text{Var}(\log(\widehat{RR}_i)$, based on the terms in the table, and give the form of the resulting $100(1-\alpha)\%$ confidence interval for RR_i. [4]

For the remainder of this part of the question you are asked to consider a method for constructing a Mantel-Haenszel (type) estimate of the pooled relative risk from K independent tables.[1] You may assume that the *unconditional* likelihood of the i-th table is given by

$$L_i = p_{1i}^{a_i}(1-p_{1i})^{b_i} p_{0i}^{c_i}(1-p_{0i})^{d_i}$$

so that if the relative risk $\psi = RR_i$ is common to all tables, then $\psi_i = p_{1i}/p_{0i} = \psi$ for all i.

(ii) Make a suitable substitution to obtain the likelihood for the i-th table in terms of ψ, and hence show that the overall log-likelihood (for K tables), is given by

$$\ell(\psi) = \sum_{i=1}^{K} \ell_i = \sum_{i=1}^{K} \Big\{ a_i \log(p_{0i}\psi) + b_i \log(1-p_{0i}\psi) \Big\}$$

(if we consider only those terms involving ψ).

[1] You should refer to the equivalent derivation for the odds-ratio given in Chapter 2 of the Lecture Notes.

Hence, show that the maximum likelihood estimator of ψ, is given by

$$\hat{\psi} = \frac{\sum_{i=1}^{K} a_i}{\sum_{i=1}^{K}(a_i + b_i)p_{0i}} \qquad (*)$$

What assumption was necessary to give the estimator in this form? [5]

(iii) Now substitute $\hat{p}_{01} = c_i/n_{0i} = c_i/(c_i + d_i)$, to obtain

$$\hat{\psi} = \frac{\sum_{i=1}^{K} a_i(c_i + d_i)/n_{0i}}{\sum_{i=1}^{K} c_i(a_i + b_i)/n_{0i}}$$

The estimator of ψ is more usually given in the form

$$\hat{\psi} = \widehat{RR}_{MH} = \frac{\sum_{i=1}^{K} a_i(c_i + d_i)/N_i}{\sum_{i=1}^{K} c_i(a_i + b_i)/N_i} \qquad (**)$$

Under what assumption is it reasonable to replace n_{0i} from the solution in part (c) with N_i in $(**)$? What nice properties does the Mantel-Haenzsel estimator given by $(**)$ share with the equivalent version for the odds-ratio? [3]

(b) A retrospective cohort study involving 248 children enrolled within one LEA area was undertaken to investigate a possible link between neurological impairment (measured by decreased reaction times) at age 11 and exposure to secondary ('passive') smoking in the home. Those classed as being exposed to secondary smoking were those who were identified as having been exposed from birth until they were at least seven years old. The data are summarised in the Table below.

Exposure to secondary smoking	Neurological impairment		
	Yes	No	Total
Yes	11	111	122
No	4	122	126
Total	15	233	248

(i) Estimate the relative risk of neurological impairment for children exposed to secondary smoking, and calculate a 95% confidence interval for this risk. [3]

(ii) Perform a formal test of no association between neurological impairment and secondary smoking for these data. What are your conclusions based on this analysis? [3]

(iii) It is possible that the results above are biased, since it is known that many in the exposed group suffering impairment were from urban areas, while in the unexposed group many of those who showed no signs of impairment were from rural areas. What type of bias might this indicate, and how could its effect have been controlled in the study? [2]

2. (a) A case-control study was undertaken to investigate the association between prostate cancer and the use of statins. There were 98 cases and 202 controls, the data from which are shown in the table below with information on the cumulative amount of statin taken

Statin use and cancer

Cumulative statin dose x (g)	Cases	Controls
No use	64	103
$0.0 < x \leq 9.3$	13	25
$9.3 < x \leq 21.2$	10	25
$21.2 < x \leq 34.8$	5	25
$x > 34.8$	6	24

(i) Estimate the dose-specific odds ratios, relative to the zero dose ('no use'), and give approximate 95% confidence intervals. Comment on your results. Does there appear to be a dose-response relationship between statin-use and subsequent cases of prostate cancer? [4]

(ii) Describe briefly the motivation behind the Armitage-Cochran test for linear trend. [You are not expected to derive the actual test or state its statistic]. *SAS, Proc Freq* gives the value of the test statistic as $Z = 2.8296$. How does this correspond to the usual χ^2 statistic? What do you conclude? [3]

(iii) Briefly explain the relevance of a dose-response relationship in Bradford Hill's criteria for assessing whether an association between two variables is evidence of a causal link. [2]

It is also possible to undertake a test for linear trend, and estimate the dose-specific odds ratios, under such an assumption using logistic regression. The following model is appropriate

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i \qquad i = 1, \ldots, 5.$$

where x_i is the midpoint of the i-th dose interval (i.e. takes the values 0, 4.7, 15.25, 28, 34.8). Fitting this model (in *SAS, Proc logistic*) results in the following output

Parameter Estimates

	Estimate	Std. Error
Intercept	-0.4821	0.1460
statin	-0.0312	0.0112

(iv) Comment on the above analysis in as much detail as possible. What do the parameters represent? What does the model suggest for the dose-specific odds ratios (relative to no use). Is there evidence of a linear trend? Plot (roughly) the dose-specific log-odds ratios and add the information from the fitted logistic regression. [5]

(b) In a 1-1 matched case-control study of asthma and nut allergies amongst undergraduates at a particular university, the following results were found.

Asthma and nut allergy amongst undergraduates

		CONTROLS	
		Asthma	No asthma
CASES	Nut allergy	11	39
	No nut allergy	18	103

(i) Briefly explain what is meant by a 1-1 matched case-control study. What is the total number of individuals involved in this study? [2]

(ii) Calculate the Mantel-Haenszel estimate of the odds ratio for the association between asthma and nut allergy, and calculate the 95% confidence interval. [2]

(iii) Carry out a formal test for no association between nut allergy and asthma. What do you conclude? [2]

3. (a) Suppose that in a sample of survival times for n patients, '*deaths*' are observed at the unique times $t_1, \ldots, t_k$, where for each observed time t_j, d_j deaths occur.

(i) Define the empirical distribution $\hat{F}(t)$ for a sample where there are no censored times, and hence justify that the empirical survivor function can be written

$$\hat{S}(t) = \frac{n - \sum_{j|t_j<t} d_j}{n}$$

Show that $\hat{S}(t)$ may alternatively be written as

$$\hat{S}(t) = \prod_{j=1}^{s}\left(1 - \frac{d_j}{r_j}\right), \quad t_s \leq t < t_{s+1} \qquad (*)$$

Explain carefully what the term r_j represents in this context, and describe briefly the form that this function takes. In what sense is this latter form intuitive if we think of survival as a discrete time process? [4]

(ii) Suppose that our observed survival times now include those which are right-censored. Explain briefly how such censored data arises, and explain the change in the interpretation of r_j which allows $(*)$ to still be be used as an estimator of the survival function. Under what name is this estimator better known? [3]

(b) A trial of 44 patients with chronic active hepatitis was undertaken to investigate whether there was any difference in the survival patterns between patients assigned to the drug prednisolone and those in an untreated control group.

The data are shown in the table below.

Survival times of chronic active hepatitis patients

Prednisolone		Control	
2	131*	2	41
6	140*	3	54
12	141*	4	61
54	143	7	63
56*	145*	10	71
68	146	22	127*
89	148*	28	140*
96	162*	29	146*
96	168	32	158*
125*	173*	37	167*
128*	181	40	182*

(right-) censored observations are indicated with an $*$.

(i) Calculate the estimated Kaplan-Meier survivor functions for each of the two groups, which correspond to the plot on the following page.[2] Does the plot suggest a difference in the survival probabilities of the two groups of patients? [5]

(ii) Carry out a log-rank test to test for a difference in survival between the two groups, explaining carefully, by reference to your calculations, the motivation behind this test procedure.[3] What do you conclude? [4]

Suppose that, alternatively, the test in part (ii) was constructed using a (Cox) proportional hazards regression model.

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta x_i, \qquad i = 1, \ldots, 44.$$

where x_i is an indicator variable assigning treatment group ($x_i = 0$ for control, and $x_i = 1$ for prednisolone). The following output is obtained from the fitted model

Parameter Estimates

	Estimate	Std.Error
prednisolone	-0.84	0.432

(iii) Explain briefly why an intercept term is not required in $(*)$. What does the fitted model say about the difference between the survivor probabilities? Which of the two approaches above do you prefer and why? What further analysis might you suggest for these data? [4]

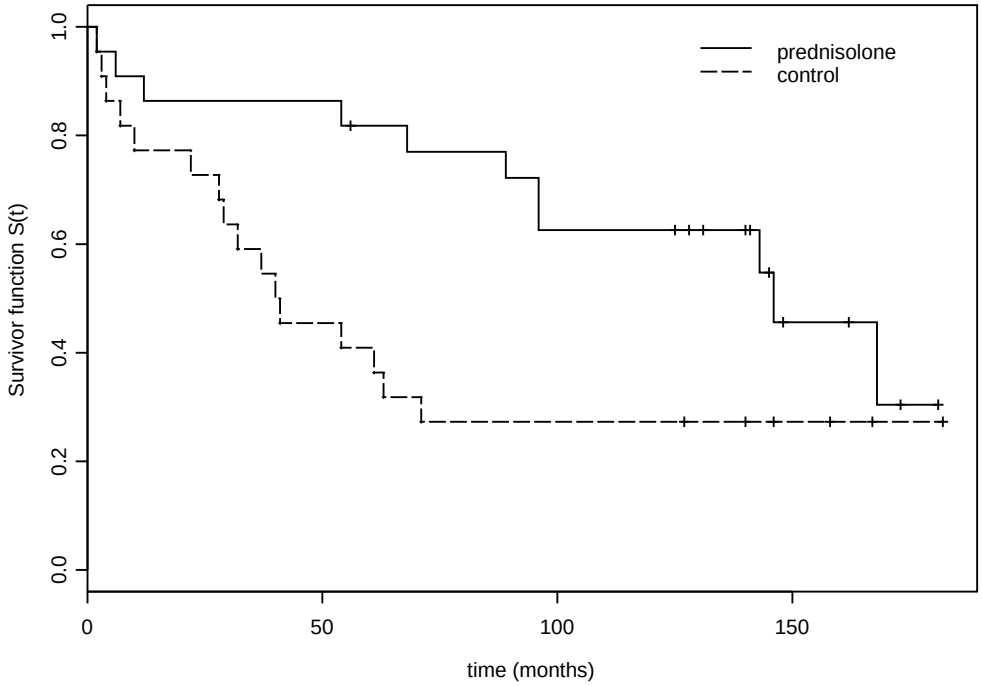
[2] Note that calculations in parts (i) and (ii) of this part of the question should be performed '*by-hand*'. You may, however, find it helpful for the purposes of the assignment to use a package such as *EXCEL*, but you should include details of your working with your solutions.

[3] Recall that in the log-rank test, the expected number of deaths in the first group at the j-th unique death time is given by

$$e_{1j} = \frac{r_{1j}d_j}{r_j} \quad \text{with variance} \quad v_{1j} = \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j^2(1 - r_j)}$$

where r_{ij} is the number at risk in the i-th group ($i = 1, 2$ at time t_j) and d_j is the corresponding number of deaths.

Estimated survivor functions of chronic active hepatitis patients



4. A meta-analysis was undertaken of some 12 studies to investigate the association between body mass index (BMI) and prognosis in breast cancer. The measured effect in each of the 12 studies chosen under a systematic review was the hazard ratio of death for those women in the largest (relative to the lowest) categories of BMI for the studies[4], with a hazard ratio greater than 1 indicating a greater probability of death at all times for women with the largest BMI.

In each of the selected studies, the estimated effect $\hat{\psi}_i$ was reported along with a corresponding 95% confidence interval for ψ, so that an appropriate meta-analysis could be carried out. The results of each of the twelve studies, including the numbers of individuals involved, and other useful quantities is included below.

Characteristics of studies for meta-analysis

Study	No. subjects	Est. hazard ratio $\hat{\psi}_i$	95% C.I.	θ_i	w_i	$w_i\hat{\theta}_i$	$w_i\hat{\theta}_i^2$	w_i^2
1	582	1.80	(0.89, 3.64)	0.588	7.744	4.552	2.676	59.971
2	838	1.40	(1.11, 1.49)	0.336	71.304	23.992	8.073	5084.313
3	213	3.89	(0.77, 19.70)	1.358	1.464	1.989	2.702	2.144
4	1170	1.70	(1.20, 2.30)	0.531	31.666	16.803	8.916	1002.714
5	1130	2.50	(1.80, 3.40)	0.916	35.598	32.618	29.888	1267.247
6	359	5.93	(1.98, 17.80)	1.780	3.193	5.683	10.116	10.193
7	1033	0.78	(0.48, 1.22)	-0.248	16.297	-4.049	1.006	265.607
8	241	0.95	(0.51, 1.78)	-0.051	9.928	-0.509	0.026	98.564
9	698	1.90	(1.00, 3.70)	0.642	9.325	5.985	3.842	86.952
10	1238	1.37	(0.99, 1.90)	0.315	36.401	11.460	3.608	1325.054
11	378	2.20	(0.90, 5.40)	0.788	4.809	3.791	2.989	23.122
12	149	0.74	(0.32, 1.71)	-0.301	5.466	-1.646	0.496	29.879
TOTAL	8029				233.196	100.669	74.336	9255.760

Assume in the first instance a fixed effects meta-analysis of the (log-) hazard ratio is appropriate, using the model

$$\hat{\theta}_i = \theta + \epsilon_i, \qquad i = 1, \ldots, 12. \tag{$*$}$$

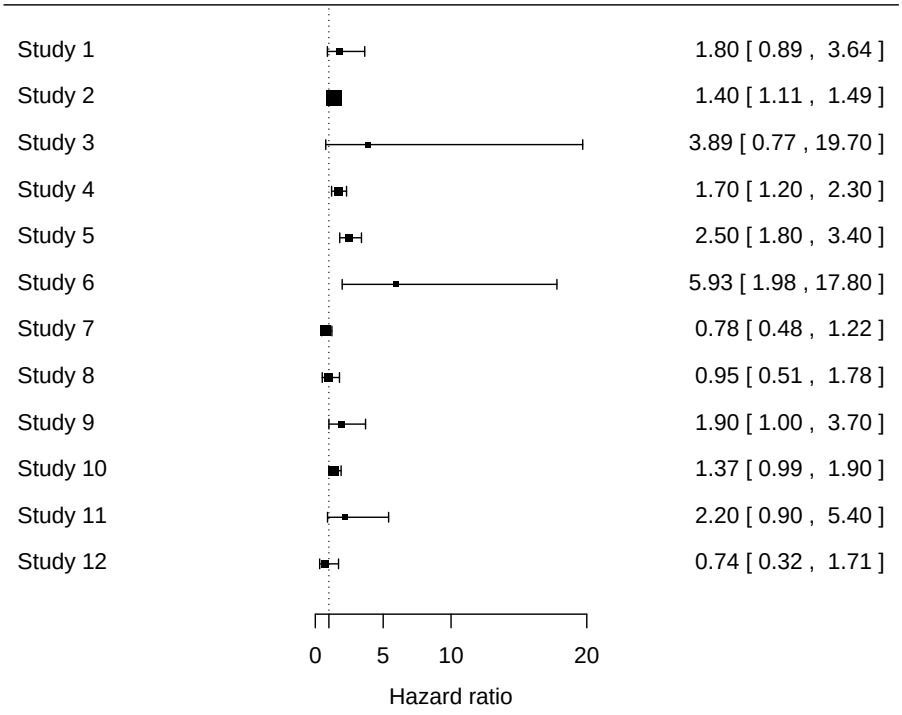
where $\epsilon_i \sim N(0, \sigma_i)$.

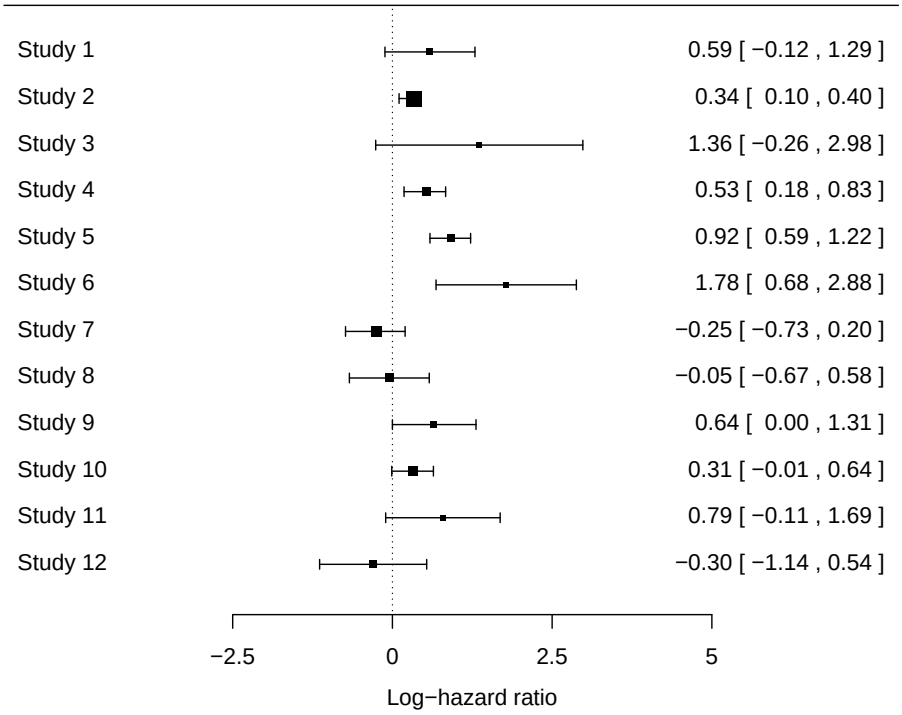
(a) What are the assumptions of the model $(*)$ that suggest it is appropriate to provide a common estimate of the log-hazard ratio, $\theta = \log(\psi)$ for prognosis from cancer and BMI from the twelve studies.

Comment on the information from the first five columns of the above table and the forest plots of the following page of the hazard and log-hazard ratios from the studies. Which studies are consistent with a higher death rate from cancer for women who have the largest BMI category, and which the lowest. [4]

[4]The hazard ratios were calculated from a Cox proportional hazards model.

Forest plots of effect size





(b) What do the weights w_i in the table represent. How are the weights in column 6 calculated from the corresponding confidence intervals. Illustrate your answer by confirming for the weight of study 1. [3]

(c) Carry out a fixed effects meta-analysis, to obtain an estimate of the common hazard ratio, $\hat{\psi}$ from the 12 studies, and calculate a 95% confidence interval for ψ.[5] [4]

(c) Is there evidence that the hazard ratio is not equal to 1? [You should carry out a formal test here]. Test the hypothesis of homogeneity between studies. [4]

You should have found evidence in part (d) of heterogeneity between studies, so that a random effects analysis may be more appropriate. i.e. model $(*)$ becomes

$$\hat{\theta}_i = \theta + b_i + \epsilon_i, \qquad i = 1, \ldots, 12 \qquad (**)$$

where $b_i \sim N(0, \tau^2)$ and $\epsilon_i \sim N(0, \sigma_i^2)$ are mutually independent.

(e) What does the term b_i in $(**)$ represent? Carry out a random effects meta analysis, where the weights are now

$$w_i^* = (w_i^{-1} + \hat{\tau}^2)^{-1}$$

with τ^2 estimated using the method of moments[6], to obtain an alternative estimate of ψ from the 12 studies, and a 95% confidence interval.

How does this estimate compare with the fixed effects estimate found in part (c)? [5]

[5] Hint: You should carry out the meta-analysis with respect to the log-hazard ratio θ and then translate your result to obtain inferences about ψ.

[6] Recall that, under the method of moments

$$\hat{\tau}^2 = \frac{Q - (r-1)}{\sum_{i=1}^r w_i - \frac{\sum_{i=1}^r w_i^2}{\sum_{i=1}^r w_i}}$$

where $Q = \sum_{i=1}^r (\hat{\theta}_i - \hat{\theta})^2$, and r is the number of studies.