Biostatistics 513 Spring 2001

Midterm Exam NAME:

This exam is to be completed without use of notes or text. Answer in the space provided.

1. [10pts] In the analysis of the HIVNET data we considered both the "nurse item", `nurse0`, and the "safety item", `q4safe0`, measured at baseline. The sample of 1000 subjects answered a number of questions regarding vaccine trial conduct in addition to these two items. We found that 617/1000 subjects correctly answered the nurse item (61.7%) while 566/1000 subjects correctly answered the safety item (56.6%). What method would you use to statistically compare the baseline knowledge of the nurse (treatment allocation) item to the baseline knowledge of the safety item? Justify your response and reference appropriate tests or estimates that you would use.

2. Graham et al. (1981) study dietary factors in the epidemiology of cancer of the larynx. Interviews were carried out with 338 male patients at Roswell Park Memorial Institute with cancer of the larynx, and with 359 male controls with diseases other than of the digestive or respiratory system (and without neoplasms).

This table compares vitamin A (IU/month) intake for the cases and the controls.

	Cases	Controls	Total
$<$50,500	98	78	176
$\geq$50,500	240	281	521
Total	338	359	697

(a) [5pts] What are appropriate null and alternative hypotheses for testing the association between vitamin A intake and cancer?

(b) [5pts] Give a point estimate for the association between vitamin A intake and cancer (ie. relative risk, odds ratio, risk difference - whichever is appropriate here) and interpret this estimate.

An additional analysis of Vitamin C showed the following results:

	Cases	Controls	Total
<1000 (mg/month) vit C	112	75	187
$\geq$ 1000 (mg/month) vit C	226	284	510
Total	338	359	697

The following statistics are available:

```
             |      Point estimate    |  [95% Conf. Interval]
             |------------------------+----------------------
Odds ratio   |         1.876578       |  1.335739    2.636309
            +-----------------------------------------------
Pearson chi-square  = 13.30;  df = 1; p-value = 0.003
```

Interpret the data comparing `low / high` vitamin C consumption by answering the following questions:

(c) [5pts] Interpret the χ^2 statistic with respect to the relationship between disease and vitamin C intake. What can we conclude from this test?

The actual data presented in Graham et al. (1981) is given as follows:

vitamin C	Cases	Controls	unadjusted OR
<1000	112	75	2.53
1000-1400	116	138	1.42
1400-1800	74	85	1.48
>1800	36	61	1.00 (reference)
Total	338	359	

```
Test of homogeneity (equal odds): chi2(3)  =    15.79
                                  Pr>chi2  =   0.0013

Score test for trend of odds:     chi2(1)  =    12.45
                                  Pr>chi2  =   0.0004
```

(d) [5pts] State (in words or in symbols that you define) the null and alternative hypotheses for testing whether there is a trend in disease status with vitamin C consumption.

H0:

H1:

Consider how a logistic regression model could be used to test for a trend in the odds of disease with increased vitamin C consumption:

(e) [5pts] Define a covariate, $X1$, representing vitamin C consumption, and define a logistic regression model using $X1$, that could be used to test for trend.

(f) [5pts] Define the null hypothesis and alternative hypotheses based on your logistic regression model that would be used to test for trend.

(g) [5pts] What test statistic would you use to execute the test of the hypothesis given in part (f) above?

(h) [5pts] Formulate a logistic regression model that could be used to estimate the disease risk associated with vitamin C consumption that adjusts for vitamin A consumption (denoted $X2$). Define your vitamin A variable X2 and give the logistic regression model.

3. The following data were taken from the manuscript: "Breast Cancer, Lactation History, and Serum Organochlorines" by Romieu et al. (2000) *AJE*. Recent studies have suggested that exposure to low levels of the toxins DDT and DDE (organochlorines) is associated with breast cancer. A case-control study of women who had given birth to at least one child was conducted in Mexico City, Mexico.

The following variables are reported in Romieu et al.:

```
DDT:     1 = 0.023-0.070 micro g / g lipids     (serum measurement)
         2 = 0.071-0.10  micro g / g lipids
         3 = 0.11-0.18   micro g / g lipids
         4 = 0.19-5.41   micro g / g lipids

DDE:     1 = 0.20-1.16  micro g / g lipids      (serum measurement)
         2 = 1.17-1.96  micro g / g lipids
         3 = 1.97-3.48  micro g / g lipids
         4 = 3.49-14.84 micro g / g lipids

POST:    0 = premenopause
         1 = postmenopause

CASE:    0 = control
         1 = case (breast cancer)

COUNT:  number of subjects
```

The goal of the study was to assess the relationship between exposure and the risk of breast cancer. A total of 126 cases were obtained and 120 community controls were also recruited.

A dichotomous exposure variable was created:

DDEhigh=1 if DDE 1.97-14.84 micro g / g lipid
DDEhigh=0 if DDE 0.20-1.96 micro g / g lipid

(a) [5pts] A crude analysis of the relationship between DDEhigh and CASE yielded:

```
                              DDEhigh=1 (high)   DDEhigh=0 (low)
      Case=1 (breast cancer)          82                 38
      Case=0 (control)                63                 63

  Odds ratio estimate:  2.158
  95% Confidence Interval for the OR:  (1.286, 3.621).
```

Interpret the odds ratio, and interpret the confidence interval for the odds ratio (is it a significant association?)

Additional analysis revealed that CASE status and menopause status (POST) were associated (OR=1.178), and menopause status was associated with exposure (OR=5.899).

A stratified analysis yielded:

```
Odds Ratios comparing CASE odds among DDEhigh=1 (high) to DDEhigh=0 (low):

    Strata     OR     95% Conf. Interval

     POST=0     1.907  (0.910, 3.997)
     POST=1     3.093  (1.257, 7.581)

Test of Homogeneity:  (Breslow-Day) chi2(1) statistic = 0.64, p-value = 0.422

                 Crude Odds Ratio Estimate:  2.158
Mantel-Haenszel Common Odds Ratio estimate:  2.326
95% Confidence Interval for Common OR:  (1.309, 4.132)
```

(b) [5pts] Is a common odds ratio estimate appropriate based on these statistics? Justify your answer.

(c) [5pts] Give an explicit interpretation of the common odds ratio estimate (OR estimate = 2.326).

(d) [5pts] If a similar stratified analysis was performed to evaluate DDEhigh but using the levels of DDT as the stratifying variable, then what would be the hypothesis of homogeneity of the odds ratios and what would be the degrees of freedom for a test of this homogeneity hypothesis?

(e) [5pts] Given the crude and adjusted analyses, would you conclude that menopause status is a confounder? Justify your answer.

A subsequent analysis used logistic regression with dummy variables to code for the variable DDE. The results of this model are:

```
Note:  DDE=1 is the reference category and no dummy variable is included

        Odds ratio  s.e.      Z     p-value     95% Conf. Interval
DDE=2    1.107       .457   0.246    0.806       0.493  2.488
DDE=3    2.213       .876   2.007    0.045       1.019  4.809
DDE=4    2.814     1.186    2.455    0.014       1.232  6.429
POST     0.796       .236  -0.770    0.441       0.445  1.423

log likelihood = -81.206
```

(f) [5pts] Interpret the odds ratio for `DDE=4`. (Describe the specific comparison that is made).

(g) [5pts] A model with only the POST variable gave a log likelihood of -170.23. Complete the following expression that refers to a likelihood ratio test comparing the model above to the null model that only has the POST variable:

Likelihood Ratio Statistic = LR = _____

Degrees of freedom for the LR Test = _____

Further analysis found that a linear model ("grouped linear") for DDE was appropriate (when compared to the dummy variable model using a LR test). Logistic regression was then used to assess whether the effect of DDE exposure appeared to depend on menopause status by fitting the model:

$$\text{logit}[\pi(X)] = -0.722 + 0.269\ \text{DDE} + -0.889\ \text{POST} + 0.242\ \text{DDE} \times \text{POST}$$

(h) [5pts] Based on this model, what is the estimated odds ratio comparing premenopausal women with DDE=3 (POST=0, DDE=3) to premenopausal women with DDE=1 (POST=0, DDE=1)?

(i) [5pts] Based on this model, what is the estimated odds ratio comparing postmenopausal women with DDE=3 (POST=1, DDE=3) to postmenopausal women with DDE=1 (POST=1, DDE=1)?

(j) [5pts] Likelihood ratio testing indicated that the DDE × POST interaction was not significant. However, additional interest is in the effect of DDE adjusting for <u>both</u> POST and DDT. What logistic regression model could be used for this question? What is(are) the parameter(s) in your model that would describe the effect of interest (adjusted DDE)?