STAT 3345: Environmental Risk Assessment

Fall Semester 2005

Instructor: J. Flemming
Office: Chase 122
Telephone: 494-1268
E-mail: flemming@mathstat.dal.ca

Lectures: Tuesday and Thursday 11:35-12:55
Room: LSC 202
Office Hours: Tuesday and Thursday 1:30-2:30

Text: Statistics For Environmental Biology and Toxicology
Walter W. Piegorsch and A. John Bailer, Chapman and Hall, 1998

Course Description: This course provides an introduction to a variety of statistical methods used in modern environmental research. We extend beyond the basic approaches like linear regression, to nonnormal data and complex and nonstandard models. The Splus Statistical Software package is used throughout the course.

Method of Evaluation:
Assignments – 25%
Midterm Exam 25%
Final Exam 50% (Note: There will be no supplemental exam)

Grading Scheme:

A+	90 and above
A	85 up to but not including 90
A-	80 up to but not including 85
B+	75 up to but not including 80
B	70 up to but not including 75
B-	65 up to but not including 70
C+	60 up to but not including 65
C	55 up to but not including 60
C-	50 up to but not including 55
D	47 up to but not including 50
F	below 47

Assignments: There will be 9 weekly assignments that will provide students the opportunity to use the techniques discussed in class to analyze data sets. Assignments will require computations, some to be done by hand (or calculator) and others using Splus. Late assignments will not be accepted.

Splus Software: There are two principal ways to access Splus:

1. Access Splus in the Chase Computer Lab (Room 007).
2. Install Splus 7 for Windows on your own computer. Details in class.

Week	Dates	Topics	Assigment*
1	September 8	Introduction and Chapter 1	
2	September 13 and 15	Chapters 2 and 3	1
3	September 20 and 22	Chapter 4	2
4	September 27 and 29	Sections 5.1 – 5.5	3
5	October 4 and 6	Sections 6.1 – 6.6	4
6	October 11 and 13	Section 7.1 and Midterm	
7	October 18 and 20	Sections 7.1 – 7.3	
8	October 25 and 27	Chapter 8	5
9	November 1 and 3	Sections 9.1 – 9.2	6
10	November 8 and 10	Sections 9.3 – 9.4	7
11	November 15 and 17	Section 11.1	8
12	November 22 and 24	Section 11.2	9
13	November 29 and December 1	Section 11.3 and Review	

*Assignments due on Thursday of week indicated

DALHOUSIE UNIVERSITY FACULTY OF SCIENCE
Department of Mathematics and Statistics
STATISTICS 3345 Final Examination

Date and Time: 12:00-3:00pm, Wednesday,

December 14, 2005

Print Name: _____
Student ID: _____
Signature : _____

The entire exam is worth 50 points. It is closed book, but you may refer to two 8.5 × 11in. sheets (both sides) of notes. The exam consists of 11 problems, the number of points allocated to each part of a problem is indicated. You may use a calculator.

1. Suppose a laboratory experiment is conducted to examine the ability of an environmental toxin to induce cancer in female mice, where the mice are exposed chronically (five times per week) to the same dose of the toxin over a 104-week period.

   (a) *1 Point* What is the population to which inferences are directed in this study?

   (b) *1 Point* What is the treatment?

   (c) *1 Point* Describe an appropriate control group.

   (d) *1 Point* In repeatedly observing the mice, week after week, what type of data is generated?

2. In an experiment to compare the effect of storage time on DNA degradation in seeds of a herbaceous plant, the frequency of chromosome aberrations in seeds stored at 22 degrees Celsius for six years was compared with that of a control group of unstored seeds. The data were recorded as the proportions of randomly selected seed cells exhibiting chromosome aberrations, out of 400 cells stored per group. Suppose this produced an observed control proportion of 6/400 aberrant cells, and a storage-group proportion of 11/400 aberrant cells.

   (a) *3 Points* Calculate the test-statistic to test the null hypothesis of no difference between underlying control and storage response probabilities, versus an alternative that the storage induced an increase in the response probability.

(b) *1 Point* For the test-statistic you have chosen in part (a), what distributional model has been assumed for the data?

(c) *1 Point* Suppose the p-value for the test-statistic calculated in part (a) is .11. If we take $\alpha = .10$ what conclusions can you draw?

(d) *1 Point* If we fail to reject H_0 when, in fact, it is false, what type of error are we making?

3. Yanagawa *et al.* (1994) report data on liver tissue damage after rats were exposed to various doses of an environmental toxin. The proportions of animals in which no liver damage was evidenced, stratified by dose, are given as follows:

Dose	Number with no damage	Total number of animals
0	16	17
0.5	15	19
2	10	19
10	11	22

Wald test statistics were calculated to assess whether any significant difference is observed between the dose groups and the control. The values obtained were $z_{10} = 1.385$, $z_{20} = 3.242$ and $z_{30} = 3.646$ along with the estimated correlations $r_{12} = .2322$, $r_{13} = .2458$ and $r_{23} = .2104$.

(a) *2 Points* We can use the Splus **qmvt** function to calculate the critical value in this case. What value would we use for the **rho** argument to this function?

(b) *2 Points* Suppose that the critical value obtained in part (a) is 2.384. What can we conclude?

(c) *1 Point* Suppose we elect **not** to use the MCC approach outlined above, and instead perform 3 separate analyses corresponding to the 3 different treatment-vs-control comparisons. What happens to the overall Type I error rate?

4. In a developmental toxicity study, the US National Toxicology Program examined *in utero* lethality in rat fetuses after maternal exposure to the narcotic analgesic codeine over a series of different doses. Results of assessing overdispersion in this data set are provided below (Note that the last row contains the p-values for each response level):

Control	Low dose: 75	Middle dose: 150	High dose: 300
r_0=8	r_1=19	r_2=16	r_3=15
$\sum Y_{0j} = 9$	$\sum Y_{1j} = 17$	$\sum Y_{2j} = 12$	$\sum Y_{3j} = 19$
$\sum N_{0j} = 101$	$\sum N_{1j} = 225$	$\sum N_{2j} = 196$	$\sum N_{3j} = 175$
$\hat{p}_0$=.089	$\hat{p}_1$=.076	$\hat{p}_2$=.061	$\hat{p}_3$=.109
T_0^2=.052	T_1^2=1.059	T_2^2=.074	T_3^2=.030
.820	.303	.785	.862

(a) *2 Points* Calculate the test statistic that is used to assess excess variability across all the exposure levels.

(b) *1 Point* Assume that the p-value corresponding to your test statistic from part (a) is .876. What can you conclude?

5. Suppose that for a particular data set of interest we wish to compare a linear model ($\mu_i = \beta_0 + \beta_1(x_i - \bar{x})$) with a quadratic model ($\mu_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2$). Note that we have assumed the responses Y_i are independently distributed as $N(\mu_i, \sigma^2)$.

   (a) *2 Points* State the null and alternative hypothesis of interest.

   (b) *1 Point* What formal test statistic would you calculate in this case?

   (c) *1 Point* When would you reject H_0?

6. (a) *2 Points* State two defining characteristics of longitudinal data.

   (b) *1 Point* How does survival data differ from longitudinal data?

   (c) *2 Points* Given an example of a setting in which censored survival data occur? Is it right-censored or left-censored?

7. Survival times, following liver transplant surgery, are shown below for six patients at the Stichemup Community Hospital.

Patient Number	Time to death (months)
1	3
2	12
3	19
4	27
5	56
6	87

   (a) *3 Points* Estimate the survival function, $S(t)$, using the Kaplan-Meier Method for each survival time given.

   (b) *2 Points* Estimate a 95% Confidence Interval for the value of $S(t)$ at 19 months.

8. In a study to compare the effect of dissolved oxygen (O_2) concentrations on survival in speckled trout, fish were exposed to one of two different O_2 concentrations, 0.94 and 1.16 mg/l. The survival times (in minutes) were:

at .94 mg/L:	20	22	24	26	26	29	29	31	34	34	41
at 1.16 mg/L:	30	30	35	35	40	45	50	58	62	70	100

Suppose a log-rank test is performed using Splus as shown below:

```
d1<-c(1,1,1,1,1,1,1,0,0,1,1,1,0,0,0,1,0,0,0,0,0,0)
n1<-c(11,10,9,8,7,6,5,4,4,4,3,2,1,1,1,1,0,0,0,0,0,0)
d2<-c(0,0,0,0,0,0,0,1,1,0,0,0,1,1,1,0,1,1,1,1,1,1)
dplus<-d1+d2
nplus<-n1+n2
```

```
e1<-(n1*dplus)/nplus
ones<-rep(1,length(d1))
num<-sum(d1-e1)
denom1<-nplus-dplus
denom2<-nplus-ones
denom3<-denom2*(nplus^2)
denom4<-n1*n2*dplus*denom1
denom5<-denom4/denom3
denom6<-denom5[1:21]
denom<-sum(denom6)
Tsqcalc<-(num^2)/denom
```

(a) *1 Point* The Splus command to define $n2$ is missing above. What is the command?

(b) *1 Point* What Splus command should we used to calculate the p-value corresponding to $Tsqcalc$ above?

(c) *2 Points* Suppose that the p-value obtained using the command of part (b) is .0002. What can we conclude? Be sure to state your conclusions with the motivation of the log-rank test in mind.

9. In order to fit a Cox proportional hazards model to the data described in Question 8, using Splus, we proceed as follows:

```
oxydata<-list(time=c(20,22,24,26,26,29,29,30,30,31,34,
34,35,40,41,45,50,58,62,70,100),status=rep(1,22),
x=c(0,0,0,0,0,0,0,1,1,0,0,0,1,1,1,0,1,1,1,1,1,1))
summary(coxph<-(Surv(time,status) ~ x, data=oxydata))
```

The resulting output is provided below:

	coef	exp(coef)	se(coef)	z	p
x	-1.86	0.156	0.569	-3.27	.0011

(a) *2 Points* Take $\alpha = .05$. What can you conclude about survival patterns?

(b) *1 Point* Suppose that a data recording error had been made and that the 100 minute survival time actually corresponded to the 0.94mg/l concentration. What change(s) would have to be made to the Splus code shown above?

(c) *1 Point* In general, what do regression mdoels for survival data allow us to incorporate into our analyses?

10. *5 Points* Which of the following statements are true?

(a) Environmental data are often subject to sources of excess variability that disturb the basic distributions for discrete data.

(b) If we use the Cochran-Armitage test when there is overdispersion it will overestimate the variability in the data.

(c) One approach to dealing with extra-Poisson variability is to assume the data has a beta-binomial distribution.

(d) The Pearson χ^2 test can be used with a contingency table to assess whether there is interaction between categories.

(e) The exponential family of models includes both the normal and binomial.

11. We are interested in studying aquatic toxicity in *Amphiascus tenuiremis* to the pesticide azinphosmethyl. The marker of aquatic toxicity was the capacity of *A. tenuiremis* to produce young that grow to the adult stage. We wish to determine if the sex of the offspring is independent of dose when considering toxicity-moderated maturation. The offspring counts of those that survived to adulthood are provided below in the form of a 3×2 contingency table:

Dose	Female	Male
Control	363	103
10 $\mu g/l$ azinphosmethyl	157	99
40 $\mu g/l$ azinphosmethyl	219	83

(a) *3 Points* Perform a test for homogeniety/independence with the data given above.

(b) *1 Point* What are the appropriate degrees of freedom for your test statistic?

(c) *1 Point* Suppose the approximate p-value is .001. What can you conclude?