

**SURVIVAL AND GROWTH CURVE ANALYSES APPLIED TO A  
BARNACLE DATA SET**

By

VIVIEN FREUND

B.A., The University of Durham, 1981

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES  
(Institute of Applied Mathematics and Statistics)

We accept this thesis as conforming  
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

December 1983

© Vivien Freund, 1983

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of

Mathematics

The University of British Columbia  
1956 Main Mall  
Vancouver, Canada  
V6T 1Y3

Date 16<sup>th</sup> December 1983

# **ABSTRACT**

A data set relating to the survival and growth of barnacles is examined. Survival distributions are compared by means of three nonparametric tests. The exponential model is then fitted for the survival distributions and a random effects model is developed for the slope. Polynomial growth curves are fitted and various hypotheses relating to the parameters are tested using firstly the model of Pothoff and Roy and secondly the model of Rao. Owing to the nature of the growth data, which is not longitudinal, this presents various statistical problems which are discussed.

A/ John Petkau

Thesis Supervisor

## TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT.....	ii
TABLE OF CONTENTS.....	iii
LIST OF TABLES.....	iv
LIST OF FIGURES.....	v
ACKNOWLEDGEMENTS.....	vi
1. INTRODUCTION .....	1
1.1 The Experiment.....	3
1.2 The Data.....	5
2. EXAMINATION OF THE SURVIVAL DATA .....	6
2.1 Description of the Survival Data.....	8
2.2 Exponential Models.....	19
2.3 Inference on the $\beta$ 's.....	31
2.4 Random Effects Model.....	36
2.5 Incorporating Netting and Dowling Lines into the Model..	46
3. EXAMINATION OF THE GROWTH DATA .....	52
3.1 Growth Data.....	52
3.2 Growth Curves for Items.....	54
3.3 The Growth Curve Models.....	56
3.4 Application of the Models.....	64
3.5 System One Oyster Lines.....	67
3.6 System Two Oyster Lines.....	76
3.7 Application of the Model to all Seventeen Lines.....	83
3.8 Extension of the Model to Allow Different Covariance Matrices for Different Items.....	91
3.9 Collapse Over Lines Within each System-type Combination.	95
3.10 Missing Data.....	98
3.11 Growth Curves for Individual Barnacles.....	101
3.12 Discussion.....	107
CONCLUSION.....	114
BIBLIOGRAPHY.....	116

LIST OF TABLES

	<u>Page</u>
Table 1    Survival Data.....	7
Table 2    Results of Fitting the Exponential Model Starting at $t = 0$ .....	23
Table 3    Results of Fitting the Exponential Model Starting at $t = 5$ .....	25

LIST OF FIGURES

	<u>Page</u>
Figure 1    Growth Curves for the System One Oyster Lines.....	73
Figure 2    Growth Curves for the System Two Oyster Lines.....	73
Figure 3    Growth Curves for the Dowling Lines.....	87
Figure 4    Growth Curves for the Netting Lines.....	87

### ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. John Petkau for all the time he spent in helping to produce this thesis. I am indebted to Dr. Nancy Reid for her sympathy and encouragement along the way, and for reading the thesis. Many thanks also to Hélène Crépeau who always had time to help.

The financial support of the Canadian Commonwealth Scholarship and Fellowship Committee is gratefully acknowledged.

## 1. INTRODUCTION

The examination of a particular data set leads, during the course of this thesis, to the discussion and application of a number of statistical techniques. These techniques fall under the two general headings of survival analysis and growth curves. The data set which was investigated consisted of survival and growth data for barnacles.

Barnacles are little creatures which live in the ocean. The first part of their lives is spent looking for a place to live, which will be some object in the ocean such as a rock, a boat or a whale. They usually choose a place where there are other barnacles. It is important that they choose a place with strong currents to bring in plankton for food. Once it has chosen a rock, or some other object, the barnacle attaches itself first using a relatively mild glue but subsequently using adult glue which is extremely strong. The barnacle can then never again move. The glue which they use is so strong that, for example, it is easier to chip the rock that the barnacle is attached to, than to pull off the barnacle. It has to be strong because of the forces that the barnacles have to stand up to in the ocean surf. If the barnacle chooses a bad place, where there is not much food, it will starve to death. In any case it might be eaten by predators.

The adult barnacle looks a little like a clam covered with a protective shell and with a "stalk" or "neck" with which it attaches itself to the object. The stalk serves to keep the rest of the barnacle away from the rock. Organs which look a little like feet, protrude from



the bottom of the shell. It is with these that the barnacle eats. The adult barnacle is approximately 4-5 cm long. Mr. Harry Goldberg, for his master's thesis in bioresource engineering at U.B.C. conducted an experiment relating to the survival and growth of barnacles. The barnacles were attached to lines of three different types of material. These lines were arranged in two systems. Barnacles were dying due to lack of food and because of predators. At each of a number of times data was collected on these barnacles, namely the number dead on each line, shell length, neck length, shell weight and neck weight. Mr. Goldberg was interested in modelling the survival distribution and in comparing the distributions for the three different types of line and for the two systems. He was also interested in fitting growth curves for each line for shell length and neck length and in comparing the curves obtained across systems and material types. Finally he was interested in the length-weight relationship for shell and neck. From our point of view the statistical methodology is of more interest than this particular data set, so the length-weight relationship, which involves straightforward regression, will not be discussed. Also growth curves will be fitted only for shell length, not for neck length. Everything that is done for shell length could be done in an identical way for neck length.

Due to the limited nature of the growth data - longitudinal data was not collected - it was not straightforward to fit meaningful growth curves and this presented interesting statistical problems.

The remainder of Section 1 is devoted to a fuller description of the experiment and the data. In Section 2 the survival data is

examined: firstly the survival distributions for various different lines are compared using nonparametric tests. Then the exponential model is fitted for each line and the slopes of the resulting curves are compared by means of the likelihood ratio test. Finally random effects models are developed for the slopes of the curves. In Section 3 growth curves are fitted for shell-length for each of the seventeen lines and various hypotheses about the parameters are tested. In this way it is determined whether system and material type are important factors in determining growth characteristics. However because of the limited data available, strong assumptions have to be made in order to be able to apply the growth curve models. Within each of the seventeen lines, there are a number of "items" (which will be described later). Growth curves are then fitted separately for each item within each line. An attempt is made to test for differences between the items within a line, but this is found to be impossible with the data available.

### 1.1 The Experiment

Seventeen lines of three different types of material were constructed as follows:

- i. Oyster lines - ropes, each with 10 or 11 oyster shells attached.
- ii. Dowling lines - ropes each with 20 pieces of wood (dowling) attached.

- iii. Netting lines - long cylindrical pieces of netting each with 10 compartments and a piece of hard rubber in each compartment.

The seventeen lines were arranged in two systems, the number of lines in each system being given below:

	Oyster	Dowling	Netting
System 1	6	2	1
System 2	6	1	1

The lines were taken out to the ocean, where barnacles became attached to the items on each line. The lines were later retrieved and set up along the coastline in a natural environment. At this stage no new barnacles could become attached and those already attached could not move. In this environment the barnacles were dying due to lack of food and because of predators. A record was made of the number of barnacles initially (at  $t=0$ ) on each line. There were approximately 200 barnacles initially on each system one oyster line, approximately 500 on each system two oyster line, approximately 1000 on each dowling line and on the system one netting line, and 1600 on the system two netting line. The actual numbers are given in Table 1. At a number of subsequent times data was collected on the barnacles.

## 1.2 The Data

### Survival Data

At  $t = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14$  and 17 weeks the number of barnacles found dead on each line was recorded. The dead barnacles were then removed. So for example the barnacles found dead at  $t = 7$  weeks were those that had died between  $t = 6$  and  $t = 7$  weeks. No barnacles died before  $t = 5$  weeks. More detailed information, namely the number of deaths at each time on each item within a line, was not available to us although it was probably recorded at some stage during the experiment. The survival data is given in Table 1.

### Growth Data

At each of the above times a haphazard sample of live barnacles was selected from each item on each line. It was assumed that within each item each barnacle had equal probability of being sampled. For each sampled barnacle the following measurements were taken:

- (i) neck length
- (ii) shell length

The data was recorded in centimetres. These barnacles were not removed from the items. The sample size was 5 except in the cases of samples from items on netting lines where it was 10. So the total number of barnacles measured at each time was:

From oyster lines:	$2 \times 6 \times 10$ (or 11) $\times 5 \approx 600$
From dowling lines:	$3 \times 20 \times 5 = 300$
From netting lines:	$2 \times 10 \times 10 = 200$

So there is an enormous amount of growth data. The growth data and the ways in which it is incomplete, are discussed more fully in Section 3.

## 2. EXAMINATION OF THE SURVIVAL DATA

In order to obtain the length-weight data, approximately 40 barnacles were removed from each type of material at each of a number of times (on a total of five occasions). In all analyses relating to the survival data, this fact was neglected. It was assumed that the removal of these barnacles would not significantly affect the analyses. This assumption seemed reasonable since the barnacles removed were chosen at random and the number removed was very small compared to the total number: in total approximately 200 barnacles were removed from each type of material, whereas the total number of barnacles initially on respectively oyster, dowling and netting lines was 6561, 2969 and 2552. The total number of barnacles on each of these types of line at the end of the study was respectively 1668, 1200 and 1121.

The initial number of barnacles on each line and the number found dead at each time are given in Table 1. No barnacles died before  $t = 5$  so the first time at which barnacles were found dead was  $t = 6$ . More detailed information, namely the number of deaths on each item within each line was not available to us.

The first question of interest was whether the survival distributions for lines of the same type within the same system were essentially the same. If this were found to be the case it would be reasonable to collapse across the six oyster lines within each system and across the two dowling lines in system 1. Then the situation would be simpler - we would have three "treatments" (types of material) within

Table 1 - Survival Data

	Initial #	Number Found Dead at Week							
		6	7	8	9	10	11	14	17
01	204	31	32	33	3	6	4	7	5
02	187	27	30	22	0	8	3	11	9
03	134	22	31	15	1	3	18	1	3
04	137	27	21	19	5	6	7	5	5
05	187	34	26	29	9	6	10	4	8
06	306	40	27	25	9	7	14	11	14
D1	937	128	122	113	32	29	41	44	45
D2	817	127	168	118	37	32	16	50	37
N1	952	75	226	119	116	19	75	10	7
07	530	30	42	69	34	35	26	39	20
08	180	9	18	12	7	14	7	9	17
09	579	24	40	39	47	78	49	37	17
010	155	14	16	28	9	9	1	7	4
011	677	20	61	80	35	62	30	67	16
012	470	7	22	33	28	38	21	33	25
D3	1215	50	89	126	77	96	61	99	32
N2	1600	19	23	118	216	175	110	75	48

each of two systems instead of the present more complicated situation of lines within treatment within system. The line effect would be known to be insignificant. Three non-parametric rank tests were performed in order to determine whether differences existed between the survival distributions of various groups of lines.

## 2.1 Description of the Non-Parametric Tests

### a. Logrank (Savage) Test

This test can be derived either as a linear rank test or in the context of Cox's proportional hazards model. It will first be considered from the latter point of view:

The proportional hazards model (Cox, 1972) for failure time is specified by the hazard function  $\lambda(t; z) = \lambda_0(t) \exp(z\beta)$  where  $T$  is failure time,  $t$  is the observed value of  $T$ ,  $z(1 \times p)$  is a vector of covariates,  $\beta(p \times 1)$  is a vector of regression coefficients and  $\lambda_0(t)$  is an arbitrary unspecified base-line hazard function. Let  $U(p \times 1)$  be the vector whose  $j$ th element is given by

$$U_j(\beta) = \frac{\partial \log L}{\partial \beta_j} \quad j = 1, \dots, p,$$

where  $L$  is the likelihood for  $\beta$ .

Let  $I(p \times p)$  be the matrix whose  $(j, h)$ th element is given by

$$I_{jh}(\beta) = \frac{-\partial^2 \log L}{\partial \beta_j \partial \beta_h}, \quad j, h = 1, \dots, p.$$

Then a test of the null hypothesis  $H_0: \beta = \beta_0$  is based on the score statistic  $U(\beta_0)$ . Under  $H_0$  and other mild regularity conditions  $U(\beta_0)$  is asymptotically normal with mean 0 and covariance matrix which can be consistently estimated by  $I(\beta_0)$ . Note that is being used to denote the observed Fisher information matrix. To find  $U$  and  $I$  the likelihood must be obtained:

Let the distinct death times be  $t_{(1)} < \dots < t_{(k)}$ .

Let  $d_i$  be the number of deaths at  $t_{(i)}$ .

Let  $n_i$  be the number at risk just prior to  $t_{(i)}$ .

In the case where ties and censoring are allowed the likelihood is given by (Kalbfleisch and Prentice, 1980):

$$L = \prod_{i=1}^k \left[ \frac{\exp(s_i \beta)}{\sum_{l \in R_{d_i}(t_{(i)})} \exp(s_l \beta)} \right]$$

where:  $s_i$  is the sum of the covariates associated with the  $d_i$

$$d_i$$

failures at  $t_{(i)}$ ,  $s_i = \sum_{j=1}^{d_i} z_{ij}$ ,  $l = (l_1, \dots, l_{d_i})$ ,  $R(t_{(i)})$  is the set of

all individuals at risk at  $t_{(i)} - 0$  and  $R_{d_i}(t_{(i)})$  is the set of all

subsets of  $d_i$  items chosen from  $R(t_{(i)})$  without replacement. The

exact partial likelihood with ties arises from a discrete model

specified by

$$\frac{\lambda(t; z) dt}{1 - \lambda(t; z) dt} = \frac{\lambda_d(t) dt}{1 - \lambda_d(t) dt} \exp(z\beta)$$



This likelihood is very laborious to compute. Peto (1972) and Breslow (1974) suggested the following approximation to L:

$$L_a = \prod_{i=1}^k \exp(s_i \beta) / \left[ \sum_{1 \leq R(t_{(i)})} \exp(z_i \beta) \right]^{d_i}$$

This approximation is good provided that  $d_i/n_i$  is small,  $i=1, \dots, k$ .

This means that at each time the number of failures must be small compared to the number at risk.

The test statistic is then  $U(\beta_0)$  where  $U_j(\beta) = \frac{\partial \log L_a}{\partial \beta_j}$ ,  $j = 1, \dots, p$ .

$U(\beta_0)$  is asymptotically normal with mean 0 and covariance matrix

estimated by  $I(\beta_0)$  where  $I_{jh}(\beta) = \frac{-\partial^2 \log L_a}{\partial \beta_j \partial \beta_h}$ ,  $j, h = 1, \dots, p$ .

In particular, to test  $H_0: \beta = 0$  (that failure times are unrelated to the covariates),  $U(0)' V^{-1} U(0)$  is compared to  $X^2_{(p)}$  tables

where  $U(0) = \sum_{i=1}^k \begin{bmatrix} s'_i & -d_i & n_i^{-1} & \sum_{1 \leq R(t_{(i)})} z'_i \end{bmatrix}$  and  $V_{jh} = I_{jh}(0)$ .

In the case where the survival distributions for samples from  $p + 1$  different populations are to be compared,  $z(lxp)$  is an indicator variable:  $z_i$  consists of a one in the  $j$ th position and zeros elsewhere if the  $i$ th observation is in the  $j$ th sample. In this case testing  $H_0: \beta=0$  is equivalent to testing that all the populations have

the same survival distribution. In this case

$$U_j(0) = \sum_{i=1}^k (d_{ji} - n_{ji} d_i/n_i) \quad j = 1, \dots, p$$

and

$$V_{jh} = \sum_{i=1}^k \frac{d_i(n_i - d_i)}{n_i(n_i - 1)} n_{ji} \left( \delta_{jh} - \frac{n_{hi}}{n_i} \right) \quad j, h = 1, \dots, p$$

where  $\delta_{jh}$  is the Kronecker delta ( $\delta_{jh} = 1$  if  $j = h$  and 0 otherwise),

$n_{ji}$  = number at risk in the  $j$ th sample just prior to  $t_{(i)}$

and  $d_{ji}$  = number of deaths at time  $t_{(i)}$  in the  $j$ th sample.

This test can also be derived as a linear rank test for comparing survival distributions which differ only with respect to location (Lawless, 1982): suppose we have  $p + 1$  distributions defined by probability density functions (p.d.f.'s)  $g(y - \theta_1), \dots, g(y - \theta_p), g(y)$  where  $y = \log$  lifetime. We wish to test that  $\theta_1 = \dots = \theta_p = 0$  i.e. that all distributions are identical. Let  $z(1 \times p)$  be an indicator variable ( $z_{ij} = 1$  if the  $i$ th observation is from sample  $j$  and  $z_{ij} = 0$  otherwise). Let  $\theta = (\theta_1, \dots, \theta_p)$ . Then, given the regression vector  $z$ , the p.d.f. of  $y$  is  $f(y|z) = g(y - z\theta')$ . Under this model we wish to test  $\theta = 0$ . We first consider the case with no censoring: suppose we obtain a sample  $y_1, \dots, y_n$  from these distributions. Let  $y_{(1)} < \dots < y_{(n)}$  be the ordered observations (assumed to be distinct) and  $N_i$  be the number of observations from distribution  $i$  ( $i=1, \dots, p+1$ ).

If  $r(1 \times n)$  is the rank vector based on the  $y_i$ 's, then a test of  $\theta=0$  can be based on  $U(0)$  ( $p \times 1$ ) whose  $j$ th element is given by

$$U_j(\theta) = \frac{\partial \log p(r; \theta)}{\partial \theta_j}$$

where  $p(r; \theta)$  is the probability mass function of  $r$ .

It follows that

$$U_j(0) = \sum_{i=1}^n z_{(i)j} \alpha_i \quad j = 1, \dots, p$$

where  $z_{(i)}$  is the indicator variable associated with  $y_{(i)}$ ,  $z_{(i)j}$  is the  $j$ th element of  $z_{(i)}$  and the  $\alpha_i$ 's are scores given by

$$\alpha_i = E \left[ \frac{-g'(y_{(i)})}{g(y_{(i)})}; \theta=0 \right].$$

To define a particular rank test, the  $\alpha_i$ 's are generated by choosing a specific p.d.f.  $g(y)$ . Then if the data actually comes from this p.d.f. the test will be asymptotically fully efficient relative to the parametric procedure based on the actual values rather than the ranks of the  $y_i$ 's. If the data arises from a different p.d.f. the rank test will be more efficient. In the case of the logrank test the scores are generated by letting  $g(y) = \exp(y - e^y)$ , the extreme value distribution.

The mean and covariance matrix for  $U(0)$  can be obtained by permutation theory arguments (Lawless, 1982). The  $\alpha_i$ 's can be chosen so that  $E(U(0))=0$ . Let  $V$  be the covariance matrix of  $U(0)$ . Then since  $U(0)$  is asymptotically normal,  $H_0: \theta=0$  can be tested by comparing  $U(0)' V^{-1} U(0)$  to  $X^2_{(p)}$  tables.

This test can be extended to accommodate censoring as follows: suppose that there are  $k$  distinct observed log lifetimes and  $n-k$

censoring times. If  $z_{(i)}$  is the indicator variable associated with  $y_{(i)}$ , let  $s_{(i)}$  be the sum of these vectors for all individuals censored in  $[y_{(i)}, y_{(i+1)})$ . Then the score statistic suggested by Prentice (1978) is  $U(0)(p \times 1)$  where

$$U_j(0) = \sum_{i=1}^k (z_{(i)j} \alpha_i + s_{(i)j} a_i), \quad j=1, \dots, p.$$

So individuals whose lifetimes are censored are given different scores  $a_i$ . The scores may be chosen so that  $E[U(0)] = 0$ . In particular, taking  $g(y) = \exp(y - e^y)$  as before, and using Prentice's method of obtaining the  $\alpha_i$  and  $a_i$  leads to the logrank test via the scores

$$\alpha_i = \sum_{j=1}^i \frac{1}{n_j} - 1, \quad a_i = \sum_{j=1}^i \frac{1}{n_j}, \quad i=1, \dots, k$$

where  $n_i$  = the number at risk just prior to  $t_{(i)} = \exp(y_{(i)})$

Then we obtain as before

$$U_j(0) = - \sum_{i=1}^k (d_{ji} - n_{ji} d_i / n_i) \quad j=1, \dots, p,$$

with notation as before. The expectation of  $U(0)$  is 0. Prentice obtains a permutation variance for  $U(0)$ . Either this or the variance obtained before in the context of Cox's model can be used. Since the scores were motivated by letting  $g(y) = \exp(y - e^y)$ , the extreme value distribution, the test is asymptotically fully efficient for detecting location differences under an extreme value model for log lifetimes or equivalently, for testing equality of lifetime distributions in a proportional hazards or Lehmann family when there is no censoring or equal censoring in all samples. This test was derived under the

assumption of no ties but may be used with a small number of ties.

### b. Wilcoxon Test

The Wilcoxon (or as it is sometimes called, Prentice's generalised Wilcoxon test) can also be derived as a linear rank test of the form

$$U_j(0) = \sum_{i=1}^k (z_{(i)j} \alpha_i + s_{(i)j} a_i) \quad j=1, \dots, p$$

for testing equality of lifetime distributions. In this case the scores are defined as (Prentice, 1978):

$$\alpha_i = 1 - 2 \prod_{j=1}^i (n_j - d_j + 1) / (n_j + 1)$$

$$a_i = 1 - \prod_{j=1}^i (n_j - d_j + 1) / (n_j + 1)$$

with notation as before.

Substituting these scores in  $U_j(0)$  gives

$$U_j(0) = -\sum F_i (d_{ji} - d_i \frac{n_{ji}}{n_i}) \quad j=1, \dots, p$$

where

$$F_i = \prod_{j=1}^i \frac{(n_j - d_j + 1)}{(n_j + 1)}.$$

An estimate of the covariance matrix of  $U(0)$  (see Prentice and Marek (1979)) is  $V$  where

$$V_{jh} = \sum_{i=1}^k F_i^2 \frac{d_i (n_i - d_i)}{(n_i - 1)} \frac{n_{ji}}{n_i} \left( \delta_{jh} - \frac{n_{hi}}{n_i} \right) \quad j, h=1, \dots, p.$$

Again  $U(0)' V^{-1}U(0)$  is compared to  $X^2_{(p)}$  tables.

The scores in this case are generated by taking  $g(y) = e^y/(1 + e^y)^2$ , the logistic density. So this test is asymptotically fully efficient for detecting location shifts when the underlying distributions are logistic. Whereas the log rank test gives equal weight to all terms  $(d_{ji} - d_i n_{ji}/n_i)$ , the Wilcoxon gives more weight to earlier events than to later ones. Hence this test is good at detecting differences early on. Again it was derived under the assumption of no ties, but may be used with a small number of ties.

### c. Logrank Test for Grouped Data

In the derivation of the log rank test in the context of the proportional hazards model it was assumed that if ties were present this was because the data arose from a discrete model. It may instead be the case that the data arises from a continuous model but that it is grouped - the actual survival time is not recorded, only the interval into which it falls. This would give rise to a slightly different likelihood and thus to a slightly different test than that obtained in a, as will now be described. Censoring is assumed only to occur just prior to the end of an interval. Assuming a proportional hazards model for the continuous data, then, if  $x_i$  represents failure in the  $i$ th interval,  $[a_{i-1}, a_i)$ , the hazard contribution at  $x_i$  for covariate  $z$  is

$$1 - (1 - \lambda_i)^{\exp(z\beta)}, \text{ where } (1 - \lambda_i) = \exp \left[ - \int_{a_{i-1}}^{a_i} \lambda_0(u) du \right] \text{ and } \lambda_0(t)$$

is the base-line hazard function.

Let  $\gamma_i = \log [-\log (1-\lambda_i)]$ ,  $i=1, \dots, k$ .

Then the likelihood is

$$L(\gamma, \beta) = \sum_{i=1}^k \left( \sum_{l \in D_i} \log \{1 - \exp[-\exp(\gamma_i + z_l \beta)]\} - \sum_{l \in R_i} \exp(\gamma_i + z_l \beta) \right)$$

where  $D_i$  is the set of labels attached to individuals failing at  $x_i$

and  $R_i$  is the set of labels attached to individuals censored at  $x_i$

or observed to survive past  $x_i$ . A test of  $H_0: \beta=0$  is based on  $U(\hat{\gamma}(0), 0)$

where  $U_j(\gamma, \beta) = \frac{\partial}{\partial \beta_j} \log L$  and  $\hat{\gamma}(0)$  is the maximum likelihood estimate of

$\gamma$  at  $\beta = 0$ . Again  $U(\hat{\gamma}(0), 0)$  is asymptotically normal with mean 0 and

covariance matrix estimated by  $I(\hat{\gamma}(0), 0)$  where  $I_{jh}(\gamma, \beta) = \frac{-\partial^2 \log L}{\partial \beta_j \partial \beta_h}$ .

Suppose that there are  $p + 1$  samples. When  $z$  is an indicator variable for the samples it follows that

$$U_j(\hat{\gamma}(0), 0) = - \sum_i n_i / d_i \log(1 - d_i / n_i) (d_{ji} - d_i n_{ji} / n_i) \quad j=1, \dots, p$$

and the elements of the covariance matrix are

$$V_{jh} = \sum_i \left( \frac{q_i}{n_i} \right) n_{ji} (\delta_{jh} - n_{hi} / n_i) \quad j, h=1, \dots, p$$

where 
$$q_i = \frac{n_i(n_i - d_i)}{d_i} \left( \log(1 - \frac{d_i}{n_i}) \right)^2$$

and other notation is as before. The test statistic is

$U(\hat{\gamma}(0), 0)' V^{-1} U(\hat{\gamma}(0), 0)$  which can be compared to  $X_{(p)}^2$  tables.

If  $d_i / n_i$  is small then  $\frac{n_i}{d_i} \log(1 - \frac{d_i}{n_i}) \approx -1$  and the test is

approximately the same as the logrank test in a.

### Application of the Tests

In our case we have continuous data which has been grouped - the actual survival times are not recorded, only the interval into which they fall. This gives rise to a large number of ties. However there are also a large number at risk and  $d_1/n_1$  is rarely bigger than 0.2 and usually much smaller.

Censoring is due only to the fact that some barnacles are still alive at the end of the study - this is type I censoring. So we have the same censoring pattern in all samples.

The underlying distribution of the survival times is not known so it is not clear whether the logrank test or the Wilcoxon test will be more powerful. We do know, however, that the Wilcoxon is more sensitive to differences early on whereas the logrank is more sensitive to differences later. If the logrank test is to be used it would be more appropriate to use the test described in c. as we have grouped data. The logrank test a. on the other hand assumes a discrete model. Also the approximation to the likelihood used in the derivation of this test assumed small values of  $d_1/n_1$ . It is not clear whether our values of  $d_1/n_1$  are sufficiently small.

The Wilcoxon was derived under the assumption of no ties and is only an approximation when, as in our case, a large number of ties are present.



Each test was applied to various groups of lines to test the equality of the survival distributions of the lines, with the following results:

Lines Compared	Logrank	Wilcoxon	Logrank for grouped data
$0_1, 0_2, 0_3, 0_4, 0_5, 0_6$	$\overset{2}{X}_5 = 34.7$	$\overset{2}{X}_5 = 31.8$	$\overset{2}{X}_5 = 35.9$
$0_7, 0_8, 0_9, 0_{10}, 0_{11}, 0_{12}$	$\overset{2}{X}_5 = 28.9$	$\overset{2}{X}_5 = 36.6$	$\overset{2}{X}_5 = 33.6$
$D_1, D_2$	$\overset{2}{X}_1 = 32.5$	$\overset{2}{X}_1 = 29.3$	$\overset{2}{X}_1 = 32.9$
$D_1, D_2, D_3$	$\overset{2}{X}_2 = 129.8$	$\overset{2}{X}_2 = 164.2$	$\overset{2}{X}_2 = 126.7$
$N_1, N_2$	$\overset{2}{X}_1 = 129.8$	$\overset{2}{X}_1 = 301.6$	$\overset{2}{X}_1 = 196.2$
$D_3, N_2$	$\overset{2}{X}_1 = 5.8$	$\overset{2}{X}_1 = 12.6$	$\overset{2}{X}_1 = 5.1$

All p-values are  $< .001$  except that obtained in the comparison of  $D_3$  and  $N_2$ . Here the logrank test for grouped data gives  $p \approx .025$ , the logrank gives  $p \approx .017$  and the Wilcoxon gives  $p < .001$ . For each group of lines, the hypothesis that their survival distributions are the same is strongly rejected. The p values are very small possibly due to the large amount of data. The smallest p-values occur when lines from different systems, for example  $N_1$  and  $N_2$ , are compared. On the other hand the survival distributions for  $D_2$  and  $N_3$ , lines of different types within the same system, appear to be much more similar judging by the

relatively small chi-squared value. This suggests that system may be a more important factor than material type in determining the survival distribution. In any case differences exist even between lines of the same type within a system and it is clearly not reasonable to collapse over these groups.

The discrepancy between the results of the two logrank tests is surprisingly small except in the comparison of  $N_1$  and  $N_2$ . This could be because  $d_1/n_1$  is always quite small in spite of the large number of ties. For the first three comparisons the Wilcoxon gives a similar result to the logrank tests but for the last three comparisons it gives a much smaller p-value than the other two tests. This could reflect the fact that the Wilcoxon is better at detecting differences early on, which is where the biggest differences lie in these cases. For example the percentage dying between weeks 6 and 7 on line  $N_1$  is 23.7 whereas on  $N_2$  this percentage is 1.4. Later on the discrepancies are not as great.

## 2.2 Exponential Models

Having compared the survival distributions for different lines nonparametrically, it was then of interest to look for a parametric model to fit the survival curves. There were two reasons for doing this. Firstly the shape of the survival curves was of interest in itself and secondly under a parametric model the task of comparing survival distributions for different lines would be simpler - instead of having to compare a large amount of data, namely number of deaths and

number at risk at each time, we would just have to compare a small number of parameters, for example slope.

With so much data it was likely that any parametric model would be rejected by a goodness-of-fit test. This shouldn't matter provided that the model captures the most important features of the data.

Because of its simplicity the exponential was the first choice of model. The exponential was first fitted to each of the 17 lines starting at  $t=0$ . So the following model was assumed:

$$P(T > t) = \begin{cases} e^{-\beta t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where  $T$  = lifetime of barnacles.

For each of the 17 lines the maximum likelihood estimate of  $\beta$  was obtained. Notation is defined as follows:



Under this model the likelihood is proportional to

$$L = (1 - e^{-\beta t_1})^{d_1} (e^{-\beta t_1} - e^{-\beta t_2})^{d_2} \dots (e^{-\beta t_8} - e^{-\beta t_9})^{d_9} (e^{-\beta t_9})^{N-Sd_1}$$

Since there were no deaths prior to  $t_1 = 5$  weeks on any of the lines,  $d_1 = 0$  and the first term is identically one. The last term is the contribution of the censored observations: under our model the probability that a barnacle survives at least until the end of the study

at  $t_9 = 17$  is  $e^{-\beta t_9}$ . The number that do survive this long is  $N - \sum d_i$  where  $N$  is the initial number on the line. To obtain the maximum likelihood estimate of  $\beta$ , note that

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=2}^9 \frac{d_i (-t_{i-1} e^{-\beta t_{i-1}} + t_i e^{-\beta t_i})}{(e^{-\beta t_{i-1}} - e^{-\beta t_i})} - t_9 (N - \sum d_i)$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = \sum_{i=2}^9 d_i \left\{ -e^{-\beta(t_i + t_{i-1})} \frac{(t_{i-1} - t_i)^2}{(e^{-\beta t_{i-1}} - e^{-\beta t_i})^2} \right\}.$$

$\frac{\partial^2 \log L}{\partial \beta^2}$  is clearly  $< 0$  for all  $\beta$  since the  $d_i$  are  $\geq 0$  (and clearly at

least one  $d_i$  is  $> 0$ ). So the turning point at the solution to  $\frac{\partial \log L}{\partial \beta} = 0$

is  $\hat{\beta}$ , the maximum likelihood estimate of  $\beta$ .

This was obtained by doing a Newton Raphson iteration:

$$\hat{\beta}_{n+1} = \hat{\beta}_n - \frac{\dot{L}(\hat{\beta}_n)}{\ddot{L}(\hat{\beta}_n)}$$

and

$$\hat{\beta}_1 = 1/\bar{T}$$

where  $\bar{T}$  = median survival time (not exact due to discreteness) and

$$\dot{L} = \frac{\partial \log L}{\partial \beta}, \quad \ddot{L} = \frac{\partial^2 \log L}{\partial \beta^2}$$

When convergence to desired accuracy is achieved  $-\ddot{L}(\hat{\beta})$  is the observed Fisher information and leads to an estimate of the variance of  $\hat{\beta}$ :

$$\text{Var } \hat{\beta} = [-\ddot{L}(\hat{\beta})]^{-1}, \quad \text{SE}(\hat{\beta}) = [\text{Var } \hat{\beta}]^{1/2}$$

After obtaining  $\hat{\beta}$ , two goodness of fit tests were carried out to check

the fit of the model for each line.

### 1. Pearson's Goodness of Fit Test

$$X^2 = \sum_{i=1}^{k+1} \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  = observed number of deaths in  $i$ th interval and

$E_i$  = expected number of deaths in  $i$ th interval

Under our model the estimate of  $E_i$  is  $\hat{E}_i = N(e^{-\hat{\beta}t_{i-1}} - e^{-\hat{\beta}t_i})$  and  $k + 1$

is the number of time intervals.  $X^2$  is compared to  $X^2_{(k-s)}$  tables

where  $s$  = number of parameters estimated in the model (=1). If  $X^2$  is

too large the fit is poor. It wasn't necessary to combine any intervals

as for all lines the expected frequency in all intervals was at least 5.

### 2. Likelihood Ratio Goodness of Fit Test

$$X^2 = 2 \sum_{i=1}^{k+1} O_i \log\left(\frac{O_i}{E_i}\right) \text{ is compared to } X^2_{(k-s)} \text{ tables where}$$

notation is as for 1.

### Results

Results are given in Table 2. Comparing to chi-squared tables with eight degrees of freedom, the p-values for the goodness-of-fit tests are all  $< .001$ . Clearly the model doesn't fit. The largest residuals were found to be at the beginning which is not surprising since the model gives a positive probability of death for  $t \in [0,5]$  but no deaths were

Table 2 - Results of Fitting Exponential Model Starting at  $t = 0$ .

	$\hat{\beta} \pm SE(\hat{\beta})$	Pearson's $X^2$	Likelihood Ratio $X^2$
$O_1$	$0.0520 \pm 0.0047$	299.3	242.7
$O_2$	$0.0506 \pm 0.0048$	221.5	192.4
$O_3$	$0.0672 \pm 0.0069$	263.7	216.8
$O_4$	$0.6540 \pm 0.0067$	187.7	167.0
$O_5$	$0.0623 \pm 0.0055$	251.9	226.2
$O_6$	$0.0374 \pm 0.0031$	232.2	215.5
$D_1$	$0.0500 \pm 0.0021$	368.1	380.4
$D_2$	$0.0680 \pm 0.0028$	94.6	106.8
$N_1$	$0.0638 \pm 0.0025$	506.4	477.0
$O_7$	$0.0444 \pm 0.0026$	183.3	154.6
$O_8$	$0.0391 \pm 0.0041$	500.1	508.6
$O_9$	$0.0455 \pm 0.0025$	218.7	238.8
$O_{10}$	$0.0475 \pm 0.0051$	905.4	844.5
$O_{11}$	$0.0432 \pm 0.0022$	1168.2	1008.0
$O_{12}$	$0.0318 \pm 0.0022$	743.9	402.8
$D_3$	$0.0402 \pm 0.0016$	1692.0	1419.0
$N_2$	$0.0367 \pm 0.0013$	1707.3	1976.0

actually observed in this period. So the same model was then fitted starting at  $t=5$ :

$$P(T > t) = \begin{cases} e^{-\beta(t-5)} & t > 5 \\ 0 & \text{otherwise} \end{cases}$$

Letting  $t' = t - 5$ , we have the following situation

$$t'_0 = 0 \quad t'_1 = 1 \quad \quad \quad t'_7 = 9 \quad t'_8 = 12$$



For this model the likelihood is

$$L = (1 - e^{-\beta t'_1})^{d_1} (e^{-\beta t'_1} - e^{-\beta t'_2})^{d_2} \dots (e^{-\beta t'_7} - e^{-\beta t'_8})^{d_8} (e^{-\beta t'_8})^{N - \sum d_i}$$

where  $t'_1 = 1, t'_2 = 2, \dots, t'_6 = 6, t'_7 = 9, t'_8 = 12$ . Now

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=2}^8 \frac{d_i (-t'_{i-1} e^{-\beta t'_{i-1}} + t'_i e^{-\beta t'_i})}{(e^{-\beta t'_{i-1}} - e^{-\beta t'_i})} + \frac{d_1 t'_1 e^{-\beta t'_1}}{(1 - e^{-\beta t'_1})} - t'_8 (N - \sum d_i).$$

The new maximum likelihood estimates of  $\beta$  were obtained and the goodness-of-fit tests repeated. Results appear in Table 3. The p-values for the tests are much larger this time but are again all  $< .001$ . The fit is much better but still poor. The model appears to fit better for the oyster lines than for the dowling or netting lines, though this might be because there is a smaller amount of data for the oyster lines.

It was possible that a better fit might be obtained by starting at some point other than  $t=5$ . So the same model was fitted again, this time

Table 3 - Results of Fitting the Exponential Model Starting at  $t = 5$

	$\hat{\beta} \pm SE(\hat{\beta})$	Pearson's $X^2$	Likelihood Ratio $X^2$
$0_1$	$0.0927 \pm 0.0084$	94.5	103.0
$0_2$	$0.0883 \pm 0.0084$	62.5	68.9
$0_3$	$0.1290 \pm 0.0130$	77.2	94.3
$0_4$	$0.1240 \pm 0.0130$	40.6	45.9
$0_5$	$0.1160 \pm 0.0100$	59.3	69.7
$0_6$	$0.0613 \pm 0.0050$	69.2	70.5
$D_1$	$0.0867 \pm 0.0037$	80.8	79.8
$D_2$	$0.1298 \pm 0.0054$	17.9	18.3
$N_1$	$0.1200 \pm 0.0047$	148.1	141.9
$0_7$	$0.0740 \pm 0.0043$	58.3	59.2
$0_8$	$0.0630 \pm 0.0065$	127.7	136.9
$0_9$	$0.0755 \pm 0.0042$	54.1	53.6
$0_{10}$	$0.0818 \pm 0.0087$	221.8	235.0
$0_{11}$	$0.0713 \pm 0.0037$	242.5	253.5
$0_{12}$	$0.0497 \pm 0.0034$	171.8	187.8
$D_3$	$0.0656 \pm 0.0026$	495.5	599.4
$N_2$	$0.0588 \pm 0.0021$	681.4	639.8



with two unknown parameters - the slope parameter  $\beta$  and the location parameter  $\alpha$  - both of which are to be determined by maximum likelihood. Since the likelihood is different for  $\alpha$  in different time intervals, to find  $\hat{\alpha}$  and  $\hat{\beta}$  it is necessary to determine a-priori in which interval the maximum likelihood estimate of  $\alpha$  lies.

Fortunately it was possible to determine a-priori that  $\hat{\alpha}$  would have to lie in the interval  $(5,6]$  assuming no deaths in  $(0,5]$  and at least one death in  $(5,6)$ . In our case these conditions are satisfied since no deaths are observed in  $(0,5]$  but on every line the first deaths are observed in  $(5,6]$ .

Proof that  $\hat{\alpha} \in (5,6]$

For  $\alpha \in (0,5]$  and no deaths in  $(0,5]$ , the likelihood is

$$L_1 = (1 - e^{-\beta(t_1 - \alpha)})^0 \prod_{i=2}^9 (e^{-\beta(t_{i-1} - \alpha)} - e^{-\beta(t_i - \alpha)})^{d_{i-1}} e^{-\beta(t_9 - \alpha)(N - \sum_{i=1}^8 d_i)}.$$



Then

$$\begin{aligned} \log L_1 &= \beta \alpha N - \beta t_9 (N - \sum_{i=1}^8 d_i) + \sum_{i=2}^9 d_{i-1} \log(e^{-\beta t_{i-1}} - e^{-\beta t_i}) \\ &= G(\beta) + \beta \alpha N \end{aligned}$$

where  $G(\beta)$  is a function of  $\beta$  independent of  $\alpha$ .

Since  $\beta > 0$ ,  $\log L_1$  is monotonic increasing in  $\alpha$  for  $\alpha \in (0, 5]$ .

So  $\hat{\alpha} \geq 5$ . Furthermore  $\hat{\alpha} \leq 6$  since  $\alpha > 6$  would mean that there would be a zero probability of death in  $(5, 6]$ , but we have assumed that deaths are observed in this interval.  $\hat{\alpha} \geq 5$  and  $\hat{\alpha} \leq 6$  together imply  $\hat{\alpha} \in (5, 6]$ .

### Determination of $\hat{\alpha}$

For  $\alpha \in (5, 6]$  the likelihood is

$$L_2 = (1 - e^{-\beta(t_2 - \alpha)})^{d_1} \prod_{i=3}^9 (e^{-\beta(t_{i-1} - \alpha)} - e^{-\beta(t_i - \alpha)})^{d_{i-1}} e^{-\beta(t_9 - \alpha)(N - \sum_{i=1}^8 d_i)}$$



Since we have already ascertained that  $\hat{\alpha} \in (5, 6]$ , this is the appropriate likelihood.

$$\frac{\partial \log L_2}{\partial \alpha} = \beta (N - d_1 / (1 - e^{-\beta(t_2 - \alpha)}))$$

$$\text{and } \frac{\partial^2 \log L_2}{\partial \alpha^2} = \frac{-\beta^2 d_1 e^{-\beta(t_2 - \alpha)}}{(1 - e^{-\beta(t_2 - \alpha)})^2}.$$

Since  $\frac{\partial^2 \log L_2}{\partial \alpha^2} < 0$  for all  $\alpha$  and  $\beta > 0$ , the solution to  $\frac{\partial \log L_2}{\partial \alpha} = 0$

yields a local maximum.  $\frac{\partial \log L_2}{\partial \alpha} = 0$  can be solved explicitly for  $\alpha$  in terms of  $\beta$ :

$$\frac{\partial \log L_2}{\partial \alpha} = 0 \implies \alpha = t_2 + \frac{1}{\beta} \log\left(\frac{N - d_1}{N}\right).$$

So if we maximise  $L_2$  over  $\alpha$  and  $\beta$  simultaneously we obtain

$$\alpha_m = t_2 + \frac{1}{\hat{\beta}} \ln\left(\frac{N - d_1}{N}\right)$$

where  $\hat{\beta}$  is the maximum likelihood estimate of  $\beta$ . Since  $t_2=6$ ,  $\hat{\beta} > 0$  and  $\log\left(\frac{N - d_1}{N}\right) < 0$ , we have  $\alpha_m < 6$ .

#### Case 1

If  $\alpha_m \in (5, 6]$ , ( $\implies \frac{1}{\hat{\beta}} \ln\left(\frac{N - d_1}{N}\right) \in (-1, 0]$ ) then  $\alpha_m$  is the

maximum likelihood estimate,  $\hat{\alpha}$ , of  $\alpha$ , since it is the solution to

$$\frac{\partial \log L_2}{\partial \alpha} = 0, \quad \frac{\partial^2 \log L_2}{\partial \alpha^2} < 0 \text{ and } L_2 \text{ is the appropriate likelihood for } \alpha$$

in this interval.

#### Case 2

If  $\alpha_m < 5$  ( $\implies \frac{1}{\hat{\beta}} \ln\left(\frac{N - d_1}{N}\right) < -1$ ), then  $\alpha_m$  is not the maximum

likelihood estimate of  $\alpha$ . It is the solution to  $\frac{\partial \log L_2}{\partial \alpha} = 0$  but  $L_2$  is

not the appropriate likelihood for  $\alpha < 5$ . In this case the true maximum likelihood estimate is  $\hat{\alpha} = 5$ .

### Proof

For given  $\beta$ ,  $\frac{\partial \log L_2}{\partial \alpha} = 0$  has a unique solution,  $\alpha_m$ , which is assumed to lie in  $(0, 5]$ . We have shown that the turning point is a maximum and therefore  $\log L_2$  is monotonic decreasing in  $\alpha$  to the right of  $\alpha_m$  and in particular for  $\hat{\alpha} \in (5, 6]$ . So  $\hat{\alpha} \leq 5$ . But we know that  $\hat{\alpha} \in (5, 6]$ . So  $\hat{\alpha} = 5$ .

### Determination of $\hat{\beta}$

$\log L_2$  was initially maximised over  $\alpha$  and  $\beta$  simultaneously to obtain  $\alpha_m$  and  $\beta_m$ . This was done via a Newton Raphson iteration as follows:

$$\text{Let } L_\alpha = \frac{\partial \log L_2}{\partial \alpha}, L_\beta = \frac{\partial \log L_2}{\partial \beta},$$

$$L_{\alpha\alpha} = \frac{\partial^2 \log L_2}{\partial \alpha^2}, L_{\alpha\beta} = \frac{\partial^2 \log L_2}{\partial \alpha \partial \beta}, L_{\beta\beta} = \frac{\partial^2 \log L_2}{\partial \beta^2}.$$

Initial estimates were given for  $\alpha$  and  $\beta$ . Then the iteration

$$(\hat{\alpha}_{n+1}, \hat{\beta}_{n+1}) = (\hat{\alpha}_n, \hat{\beta}_n) - (L_\alpha, L_\beta) \bigg|_{(\hat{\alpha}_n, \hat{\beta}_n)} \left[ \begin{array}{cc} L_{\alpha\alpha} & L_{\alpha\beta} \\ L_{\alpha\beta} & L_{\beta\beta} \end{array} \right]^{-1} \bigg|_{(\hat{\alpha}_n, \hat{\beta}_n)}$$

was used until convergence to desired accuracy was achieved.

### Case 1

Suppose we obtain  $\alpha_m \in (5, 6]$ .

Then  $\alpha_m$  and  $\beta_m$  are the true maximum likelihood estimates  $\hat{\alpha}$  and  $\hat{\beta}$ .

In this case an estimate of the covariance matrix of  $(\hat{\alpha}, \hat{\beta})$  is given by:

$$\left[ \begin{array}{cc} L_{\alpha\alpha} & L_{\alpha\beta} \\ L_{\alpha\beta} & L_{\beta\beta} \end{array} \right]_{(\hat{\alpha}, \hat{\beta})}^{-1},$$

the inverse of the observed Fisher information. In our case, since  $\frac{\partial \log L_2}{\partial \alpha} = 0$  can be solved explicitly to find  $\hat{\alpha} = \alpha(\beta)$ , we could instead have substituted for  $\alpha$  in terms of  $\beta$  in  $\frac{\partial \log L_2}{\partial \beta} = 0$ . Then we would have had just one equation in one unknown. This could have been solved by a Newton Raphson iteration to find  $\hat{\beta}$  and its standard error. Then  $\hat{\alpha} = \alpha(\hat{\beta})$  and the variance of  $\hat{\alpha}$  would have been estimated by

$$-L_{\beta\beta} (L_{\alpha\alpha} L_{\beta\beta} - L_{\alpha\beta}^2)^{-1} \Big|_{(\hat{\alpha}, \hat{\beta})}$$

## Case 2

Suppose we obtain  $\alpha_m < 5$ . Then  $\alpha_m$  and  $\beta_m$  are not the true maximum likelihood estimates. As already shown,  $\hat{\alpha} = 5$ . Knowing that  $\hat{\alpha} = 5$ , we can think of  $\alpha$  as fixed. We maximise with respect to  $\beta$  given that  $\alpha = 5$ . This means that when we fitted the exponential starting at  $t = 5$ , we already had the best possible fit as  $t = 5$  was the best point at which to start. So  $\hat{\beta}$  is as obtained in this previous analysis. As for case 1 the variance of  $\hat{\beta}$  is estimated by  $-L_{\alpha\alpha} (L_{\alpha\alpha} L_{\beta\beta} - L_{\alpha\beta}^2)^{-1} \Big|_{(\hat{\alpha}, \hat{\beta})}$

The  $X^2$  values are the same as those obtained in the previous analysis starting at  $t = 5$ . We have one degree of freedom fewer but this shouldn't affect the conclusions.

## Results

In those cases for which  $\hat{\alpha} > 5$  the fit was slightly improved by maximising over two parameters. This improvement was reflected in the slightly smaller  $X^2$  values. For seven of the seventeen lines it was found that  $t = 5$  was not the ideal place to start. The results for these lines are as follows:

	$\hat{\beta} \pm SE(\hat{\beta})$	$\hat{\alpha} \pm SE(\hat{\alpha})$	Pearson's $X^2$	Likelihood Ratio $X^2$
$O_7$	$0.0763 \pm .0047$	$5.24 \pm .147$	76.1	77.8
$O_8$	$0.0646 \pm .0071$	$5.21 \pm .279$	16.9	17.8
$O_9$	$0.0805 \pm .0046$	$5.47 \pm .112$	129.2	130.8
$O_{11}$	$0.0774 \pm .0041$	$5.61 \pm .089$	99.7	114.2
$O_{12}$	$0.0541 \pm .0038$	$5.72 \pm .107$	37.3	36.7
$N_1$	$0.1281 \pm .0054$	$5.36 \pm .079$	454.2	584.8
$N_2$	$0.0624 \pm .0019$	$5.81 \pm .152$	682.1	580.3

Values of  $\hat{\alpha}$  are generally smaller for the system one oyster lines and for the dowling lines than for other lines. It appears that barnacles on these lines start dying sooner. However the scale parameter  $\beta$ , which relates to the rate at which the barnacles die off, is of more interest.

### 2.3 Inference on the $\beta$ 's

Although the model doesn't fit very well,  $\hat{\beta}$  is a good summary statistic and contains a lot of information about the lines. So many

analyses relating to the  $\hat{\beta}$  values were then carried out. All of these analyses were based on the estimates of  $\beta$  obtained by maximising over two parameters. Previously we had compared the survival distributions of certain groups of lines nonparametrically. Now we were in a position to compare the distributions of the same groups of lines parametrically by testing whether all system one oyster lines (say) had the same underlying  $\beta$  value. This was done by means of the likelihood ratio test as follows:

To test  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6$

$2[\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}}) - \log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)]$  is compared to  $X^2_{(q)}$  tables and  $H_0$  is rejected if this statistic is too large. Here  $\beta_j$  denotes the true slope for the  $j$ th oyster line on system one. This test is valid provided that samples are large, which they are. Here  $q$  = number of constraints imposed by  $H_0$ ,  $\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}})$  = value of log likelihood at the maximum likelihood estimates and  $\log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)$  = value of log likelihood at the maximum likelihood estimates obtained under  $H_0$ . In our case  $L(\underline{\beta}, \underline{\alpha})$  is the likelihood of obtaining the observed numbers of deaths on each line if the  $j$ th line follows the exponential model with parameters  $\beta_j$  and  $\alpha_j$  which may be different for each line.

So to obtain  $\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}})$  we maximize for each line separately (which we have already done) and add the maximum values of the log likelihood.  $L(\underline{\beta}_0, \underline{\alpha}_0)$  is the likelihood of obtaining the observed numbers of deaths on each line if the  $j$ th line follows the exponential model with parameters  $\beta$  and  $\alpha_j$ . So all lines must have a common  $\beta$  but their

location parameters  $\alpha_1, \dots, \alpha_6$  may be different. This time the six parts cannot be maximized separately.

$$\begin{aligned} \log L(\underline{\beta}_0, \underline{\alpha}_0) = & \sum_{i=3}^9 d_{i-1} \log(e^{-\beta t_{i-1}} - e^{-\beta t_i}) - \beta t_9 (N - \sum_{i=1}^8 d_i) \\ & + \beta [\alpha_1(N_1 - d_{11}) + \alpha_2(N_2 - d_{12}) + \dots + \alpha_6(N_6 - d_{16})] \\ & + d_{11} \log(1 - e^{-\beta(t_2 - \alpha_1)}) + \dots + d_{16} \log(1 - e^{-\beta(t_2 - \alpha_6)}) \end{aligned}$$

where  $N_j$  = initial number of barnacles on line  $j$ ,

$d_{ij}$  = number of deaths at time  $t_i$  on line  $j$ ,

$d_i = \sum_{j=1}^6 d_{ij}$  = total number of deaths at time  $i$  and  $N = N_1 + N_2 + \dots + N_6$ .

In this case,  $\frac{\partial \log L}{\partial \alpha_j} = 0$  can be solved to give

$$\alpha_j = t_2 + \frac{1}{\beta} \log\left(\frac{N_j - d_{1j}}{N_j}\right) \quad j=1, 2, \dots, 6.$$

This is clearly a maximum since  $\frac{\partial^2 \log L}{\partial \alpha_j^2} < 0$  for all  $\alpha_j$  and  $\beta > 0$ .

This can be substituted in  $\frac{\partial \log L}{\partial \beta}$  to give

$$\frac{\partial \log L}{\partial \beta} = -t_9 (N - \sum_{i=1}^8 d_i) + \sum_{i=3}^9 \frac{d_{i-1} (t_i e^{-\beta t_i} - t_{i-1} e^{-\beta t_{i-1}})}{(e^{-\beta t_{i-1}} - e^{-\beta t_i})} + t_2 (N - d_1).$$

A Newton Raphson iteration using

$$\frac{\partial^2 \log L}{\partial \beta^2} = \sum_{i=3}^9 d_{i-1} \begin{bmatrix} -\beta(t_i + t_{i-1}) \\ -e^{-\beta(t_i + t_{i-1})} \end{bmatrix} \begin{bmatrix} t_{i-1} - t_i \\ \frac{e^{-\beta t_{i-1}} - e^{-\beta t_i}}{e^{-\beta t_{i-1}} - e^{-\beta t_i}} \end{bmatrix}^2$$



which is clearly  $< 0$  then yields  $\hat{\beta}$ . It is then straightforward to obtain  $\hat{\alpha}_j$ ,  $j=1, 2, \dots, 6$  and to substitute the maximum likelihood estimates into  $\log L(\underline{\beta}_0, \underline{\alpha}_0)$  to obtain  $\log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)$ .

### Results

1. Test  $H_0$ : that all system one oyster lines have the same  $\beta$  value.

$$2[\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}}) - \log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)] = 2(-2190 - (-2257)) = 135.$$

Refer to  $X^2_5$  tables.  $P \ll .001$ .

2. Test  $H_0$ : that all system two oyster lines have the same  $\beta$  value.

$$2[\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}}) - \log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)] = 2(-4740 - (-4753)) = 26.$$

Refer to  $X^2_5$  tables.  $P \ll .001$ .

3. Test  $H_0$ :  $\beta_{D_1} = \beta_{D_2} = \beta_{D_3}$ , that all the dowling lines have the same  $\beta$  value.

$$2[\log L(\hat{\underline{\beta}}, \hat{\underline{\alpha}}) - \log L(\hat{\underline{\beta}}_0, \hat{\underline{\alpha}}_0)] = 2(-5648 - (-5759)) = 222.$$

Refer to  $X^2_2$  tables.  $P \ll .001$ .

4. Test  $H_0$ :  $\beta_{D_1} = \beta_{D_2}$ , that the system one dowling lines have the same  $\beta$  value. Since we are comparing only two  $\beta$  values, a t-test can be used, provided that the sample from which  $\beta_{D_1}$  and  $\beta_{D_2}$  are estimated is reasonably large. In our case it is very large. The use of a t-test can be justified as follows: in general, if  $\hat{\theta}$  is a maximum likelihood estimate of  $\theta$ , then  $\hat{\theta}$  is asymptotically normal with mean  $\theta$  and

covariance matrix  $\{E[\frac{-\partial^2 \log L}{\partial \theta^2}]\}^{-1}$  which can be estimated by  $\{\frac{-\partial^2 \log L}{\partial \theta^2}\bigg|_{\hat{\theta}}\}^{-1}$

This arises from the central limit theorem applied to the sum of the contributions to the log likelihood. Since  $\hat{\beta}_{D_1}$  and  $\hat{\beta}_{D_2}$  are maximum likelihood estimates and since their variances are estimated, under  $H_0$

and for large samples, 
$$\frac{\hat{\beta}_{D_1} - \hat{\beta}_{D_2}}{\sqrt{(\hat{SE}(\hat{\beta}_{D_1}))^2 + (\hat{SE}(\hat{\beta}_{D_2}))^2}}$$
 might reasonably be

compared to t-tables. This statistic works out to 6.6 and  $H_0$  is rejected with  $p < .001$ .

5. Test  $H_0: \beta_{N_1} = \beta_{N_2}$ , that the two netting lines have the same  $\beta$  value.

$$\frac{\hat{\beta}_{N_1} - \hat{\beta}_{N_2}}{\sqrt{(\hat{SE}(\hat{\beta}_{N_1}))^2 + (\hat{SE}(\hat{\beta}_{N_2}))^2}} = 11.5$$
 is referred to a t-distribution. Again,

$p < .001$  and  $H_0$  is rejected.

Differences were found in every set of  $\beta$  values. The  $\beta$  values for the system two oyster lines have a much smaller range of values than the values for the system one oyster lines as indicated by the smaller  $X^2$  value. The non-parametric tests also found the survival distributions to be different for different lines but found the system one oyster lines and the system two oyster lines to be approximately equally homogenous in their survival distributions. This discrepancy could arise from the fact that our parametric model doesn't fit the data very well.

### Other Parametric Models

The Weibull would have been the natural choice of a model to fit next. Since this model has two parameters (or three if a location parameter is included) and is a generalization of the exponential, it is likely that we would obtain a somewhat better fit. However the Weibull was not pursued for two reasons. Firstly, with so much data, the improvement in fit would probably be slight (in fact any model would probably be rejected). Secondly, the subsequent interpretation would be difficult - it would not be clear which parameter to focus on, and meaningful results would be hard to obtain with so many parameters and

#### 2.4 Random Effects Model

Instead of looking at these particular seventeen lines, it might be of more interest to think of these lines as a sample from an infinite population of lines. So we could think of an infinite population of lines of each type (material) within each system. For each of these subpopulations there would be a true underlying  $\beta$  value. It would be of interest to compare the underlying  $\beta$  values for the different "treatments" (types of material) and for the two systems, by examining the  $\hat{\beta}$  values obtained for our sample of seventeen lines. With only one netting line in each system and only one dowing line in system two it would be difficult to obtain an estimate of line-to-line variance separately for these treatments. So initially a random effects model was developed incorporating only the oyster lines. Throughout this

section, for those lines for which  $\hat{\alpha} = 5$ , the estimate of the variance of  $\hat{\beta}$  that was used was that obtained when the exponential model starting at  $t = 5$  was fitted. Although, strictly speaking, the variance of  $\hat{\beta}$  should have been estimated instead in the manner described in Section 2.2, any resulting inaccuracies were likely to be negligible.

Let  $\beta_1$  and  $\beta_2$  be the true underlying parameters for the hypothetical population of system one and system two oyster lines respectively. We wish to test  $H_0: \beta_1 = \beta_2$ . If  $H_0$  is rejected, this would suggest that a significant system effect is present. We have a sample of six lines from each population. Let  $\beta_{ij}$  be the true  $\beta$  value for the  $j$ th line in the  $i$ th system and  $\hat{\beta}_{ij}$  be the maximum likelihood estimate of  $\beta_{ij}$ .  $V_{ij} = SE(\hat{\beta}_{ij})^2$  is assumed to be a known constant (not unreasonable considering the large amount of data with which  $\beta_{ij}$  was estimated).

Then, provided that the model fits, since  $\hat{\beta}_{ij}$  is a maximum likelihood estimate based on a large sample, it seems reasonable to assume

$$\hat{\beta}_{ij} | \beta_{ij} \sim N(\beta_{ij}, V_{ij}).$$

We assume further, that the  $\beta_{ij}$  are independently distributed and that

$$\beta_{ij} \sim N(\beta_i, \sigma_i^2)$$

where  $\sigma_i^2$  is the line-to-line variance (unknown) of the true  $\beta_{ij}$  values for the infinite population of lines in system  $i$ . It is not clear that this second assumption is entirely reasonable. We can only proceed while bearing in mind that a strong assumption is being made. It follows that

$$\hat{\beta}_{ij} \sim N(\beta_i, \sigma_i^2 + V_{ij})$$

### Naive Analysis

Initially a naive analysis was carried out in which two fairly strong assumptions were made:

1. It was assumed that the standard errors of the  $\hat{\beta}_{ij}$  were negligible compared to the line-to-line variance, i.e.,  $V_{ij} \ll \sigma_i^2$ .

It seemed reasonable to expect that the  $V_{ij}$  would be small since  $\beta_{ij}$  was estimated with a large amount of data. By ignoring  $V_{ij}$  we were in effect, treating  $\beta_{ij}$  as known exactly.

Since we are neglecting the  $V_{ij}$  an estimate of  $\sigma_i$  is provided by the sample standard deviation of the  $\hat{\beta}_{ij}$ 's. The previous results yield

$$\hat{\sigma}_1^2 = s_1^2 = \frac{1}{5} \sum_j (\hat{\beta}_{1j} - \bar{\hat{\beta}}_1)^2 \quad \text{where} \quad \bar{\hat{\beta}}_1 = \frac{1}{6} \sum_j \hat{\beta}_{1j} = 0.1019,$$

$$\hat{\sigma}_2^2 = s_2^2 = \frac{1}{5} \sum_j (\hat{\beta}_{2j} - \bar{\hat{\beta}}_2)^2 \quad \text{where} \quad \bar{\hat{\beta}}_2 = \frac{1}{6} \sum_j \hat{\beta}_{2j} = 0.0724.$$

We obtain  $\hat{\sigma}_1 = .026$  and  $\hat{\sigma}_2 = .011$ . The values of  $\sqrt{V_{ij}}$  in system one range from .0050 to .0130 and in system two from .0034 to .0087; while the  $V_{ij}$  are smaller than our estimates of the  $\sigma_i^2$ , they are not entirely negligible.

2. The second assumption was that the line-to-line variance was the same for both systems, i.e.  $\sigma_1 = \sigma_2$ . Since we have assumed that the

$\beta_{ij}$ 's are normally distributed we can test the hypothesis  $\sigma_1 = \sigma_2$  by comparing  $S_1^2/S_2^2$  to  $F_{5,5}$  tables.

$$\frac{S_1^2}{S_2^2} = 5.67 \implies p \approx .05.$$

So the assumption  $\sigma_1 = \sigma_2$  appears to be a fairly strong one.

Under assumptions 1. and 2. we have a sample of six  $\beta$  values from each of two normal populations with equal variance. So  $H_0: \beta_1 = \beta_2$  can be tested by a two sample t-test:

$$\frac{\bar{\beta}_1 - \bar{\beta}_2}{S_p \sqrt{\frac{1}{6} + \frac{1}{6}}} = 2.57 \implies p = 0.015$$

$$\text{where } S_p = \sqrt{\frac{5S_1^2 + 5S_2^2}{10}} = 0.020.$$

This seems to suggest that there is a difference between the underlying  $\beta$  values for the two systems. However this analysis relied on two fairly strong assumptions. The second assumption,  $\sigma_1 = \sigma_2$  was tested by means of an F-test, which is particularly sensitive to our assumption that the  $\beta_{ij}$ 's are normal. A more careful analysis is called for.

### Second Analysis

This time the standard error of the  $\hat{\beta}_{ij}$ 's was not neglected. This analysis was done both assuming  $\sigma_1 = \sigma_2$  and also without this assumption.

**Case 1: Assume  $\sigma_1 = \sigma_2 = \sigma$**

If we assume that the line-to-line variance is the same in both systems then there are three parameters -  $\beta_1$ ,  $\beta_2$  and  $\sigma$  to be estimated. since  $\hat{\beta}_{ij} \sim N(\beta_i, \sigma^2 + v_{ij})$ , the likelihood of obtaining the twelve  $\hat{\beta}$ 's that we obtained is

$$L = \prod_{i=1}^2 \prod_{j=1}^6 \frac{1}{\sqrt{2\pi} \sqrt{\sigma^2 + v_{ij}}} \exp \left[ -1/2 \frac{(\hat{\beta}_{ij} - \beta_i)^2}{(\sigma^2 + v_{ij})} \right].$$

Our estimate of the variance of  $\hat{\beta}_{ij}$  was made with a large amount of data, so it can be assumed to be close to the true variance  $v_{ij}$ . So we replace  $v_{ij}$  with this estimate, thus treating  $v_{ij}$  as a known constant.

$$\log L = -\frac{1}{2} \sum_i \sum_j \left[ \log(\sigma^2 + v_{ij}) + \frac{(\hat{\beta}_{ij} - \beta_i)^2}{(\sigma^2 + v_{ij})} \right] + \text{constant}$$

$$\frac{\partial \log L}{\partial \beta_1} = \sum_j \frac{(\hat{\beta}_{1j} - \beta_1)}{(\sigma^2 + v_{1j})}, \quad \frac{\partial^2 \log L}{\partial \beta_1^2} = -\sum_j \frac{1}{(\sigma^2 + v_{1j})} (< 0)$$

$$\frac{\partial \log L}{\partial \beta_2} = \sum_j \frac{(\hat{\beta}_{2j} - \beta_2)}{(\sigma^2 + v_{2j})}, \quad \frac{\partial^2 \log L}{\partial \beta_2^2} = -\sum_j \frac{1}{(\sigma^2 + v_{2j})} (< 0)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{1}{2} \sum_i \sum_j \left[ \frac{1}{(\sigma^2 + v_{ij})} - \frac{(\hat{\beta}_{ij} - \beta_i)^2}{(\sigma^2 + v_{ij})^2} \right].$$

To find the maximum likelihood estimates of  $\beta_1$ ,  $\beta_2$  and  $\sigma$  we need to solve simultaneously  $\frac{\partial \log L}{\partial \beta_1} = \frac{\partial \log L}{\partial \beta_2} = \frac{\partial \log L}{\partial \sigma^2} = 0$ .

$$\text{Let } w_{ij} = \frac{1}{\sigma^2 + v_{ij}} \quad i = 1, 2.$$

$$\text{Then } \frac{\partial \log L}{\partial \beta_i} = 0 \implies \beta_i = \frac{\sum_j w_{ij} \hat{\beta}_{ij}}{\sum_j w_{ij}} \quad i = 1, 2.$$

Substituting the above values ( $\beta_1^*$  and  $\beta_2^*$  say) in  $\frac{\partial \log L}{\partial \sigma^2}$ ,

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \implies \sum_i \sum_j [w_{ij} - w_{ij}^2 (\hat{\beta}_{ij} - \frac{\sum_j w_{ij} \hat{\beta}_{ij}}{\sum_j w_{ij}})^2] = 0 \text{ and}$$

$$\frac{\partial^2 \log L(\beta_1^*, \beta_2^*, \sigma^2)}{\partial \sigma^2^2} = \frac{1}{2} \sum_i \sum_j w_{ij}^2$$

$$+ \sum_i \sum_j w_{ij}^2 (\hat{\beta}_{ij} - \frac{\sum_j w_{ij} \hat{\beta}_{ij}}{\sum_j w_{ij}}) \{ -w_{ij} (\hat{\beta}_{ij} - \frac{\sum_j w_{ij} \hat{\beta}_{ij}}{\sum_j w_{ij}}) + (\frac{\sum_j w_{ij}^2 \hat{\beta}_{ij}}{\sum_j w_{ij}} - \frac{\sum_j w_{ij} \hat{\beta}_{ij}}{\sum_j w_{ij}} \frac{\sum_j w_{ij}^2}{\sum_j w_{ij}}) \}.$$

$\frac{\partial \log L}{\partial \sigma^2} = 0$  can be solved by a Newton Raphson iteration to obtain  $\hat{\sigma}^2$ .

A numerical check can be done to verify that  $\frac{\partial \log L}{\partial \sigma^2} = 0$  does in fact

yield a maximum. Furthermore since  $\frac{\partial^2 \log L}{\partial \beta_1^2}$  and  $\frac{\partial^2 \log L}{\partial \beta_2^2}$  are both  $< 0$ ,



$\frac{\partial \log L}{\partial \beta_1} = 0$  and  $\frac{\partial \log L}{\partial \beta_2} = 0$  will certainly yield maxima. Having obtained  $\hat{\sigma}^2$

it is straightforward to obtain  $\hat{\beta}_1$  and  $\hat{\beta}_2$  as we have explicit expressions for them in terms of  $\hat{\sigma}^2$ .

To test whether the underlying parameters  $\beta_1$  and  $\beta_2$  are the same, we need to know the variance of our estimates. An estimate of the covariance matrix of  $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$  is provided by the inverse of the negative of the matrix of second derivatives of  $\log L$  evaluated at the maximum likelihood estimates.

Alternatively we can proceed as follows:

$$\hat{\beta}_i = \frac{\sum_j \hat{w}_{ij} \hat{\beta}_{ij}}{\sum_j \hat{w}_{ij}}, \quad \text{where } \hat{w}_{ij} = \frac{1}{\hat{\sigma}^2 + v_{ij}} \quad i = 1, 2.$$

Suppose the  $w_{ij}$  were known. Then we would have

$$\hat{\beta}_i = \frac{\sum_j w_{ij} \beta_{ij}}{\sum_j w_{ij}} \quad \text{and} \quad \text{var } \hat{\beta}_i = \frac{1}{\sum_j w_{ij}} \quad i = 1, 2.$$

Substituting our estimate  $\hat{w}_{ij}$  of  $w_{ij}$  we obtain the following rough

approximation to the variance of  $\hat{\beta}_i$ :

$$\text{var } \hat{\beta}_i \approx \frac{1}{\sum_j \frac{1}{(\hat{\sigma}^2 + v_{ij})}} \quad i = 1, 2.$$

## Results

$\hat{\beta}_1 = 0.0983$ ,  $\hat{\beta}_2 = 0.0722$ ,  $\hat{\sigma} = 0.0158$ .

Estimated  $SE(\hat{\beta}_1) = 0.0076$ , estimated  $SE(\hat{\beta}_2) = 0.0068$ .

These lead to

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{(\hat{SE}(\hat{\beta}_1))^2 + (\hat{SE}(\hat{\beta}_2))^2}} = 2.56$$

which is almost identical to the value of the statistic obtained in the naive analysis. Comparing the value 2.56 to normal tables would again suggest that there is a significant difference between  $\beta_1$  and  $\beta_2$ .

However this test is not completely legitimate for two reasons:

Firstly  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are not independent since they both involve  $\hat{\sigma}^2$  and secondly it would seem more reasonable to compare the value of the statistic to t-tables as the variances of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are only estimates. The problem is that it is not clear how many degrees of freedom are appropriate. If the square of the denominator of our statistic were a linear combination of independent chi-squared variables, Satterthwaite's approximation (Satterthwaite, 1946) could be used to estimate the degrees of freedom.

However our estimates of the variances of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are of a more complicated form and Satterthwaite's approximation is not applicable here.

Also, the whole analysis was based on our assumption of normality for the  $\beta_{ij}$ 's. Because of these problems an exact p-value cannot be given. However the value of our test statistic is sufficiently large to

suggest a significant difference between  $\beta_1$  and  $\beta_2$ , even allowing for slight inaccuracies.

**Case 2:  $\sigma_1$  and  $\sigma_2$  not assumed equal**

If  $\sigma_1$  and  $\sigma_2$  are not assumed to be equal then we need estimates of four parameters -  $\beta_1, \beta_2, \sigma_1, \sigma_2$ . For the system one oyster lines the likelihood under our model of obtaining  $\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{16}$  is

$$L = \prod_j \frac{1}{\sqrt{2\pi} \sqrt{\sigma_1^2 + V_{1j}}} \exp \left[ -1/2 \frac{(\hat{\beta}_{1j} - \beta_1)^2}{(\sigma_1^2 + V_{1j})} \right]$$

and

$$\log L = -\frac{1}{2} \sum_j \left[ \ln (\sigma_1^2 + V_{1j}) + \frac{(\hat{\beta}_{1j} - \beta_1)^2}{(\sigma_1^2 + V_{1j})} \right] + \text{constant}.$$

If we let  $w_{1j} = \frac{1}{(\sigma_1^2 + V_{1j})}$ , then

$$\frac{\partial \log L}{\partial \beta_1} = 0 \implies \beta_1 = \frac{\sum w_{1j} \hat{\beta}_{1j}}{\sum w_{1j}}.$$

As before we can substitute in  $\frac{\partial \log L}{\partial \sigma_1^2}$ . Then  $\frac{\partial \log L}{\partial \sigma_1^2} = 0$  can be solved by

Newton Raphson to obtain  $\hat{\sigma}_1^2$ . Then it is easy to find  $\hat{\beta}_1$ . Analogously to the previous case, an estimate of the variance of  $\hat{\beta}_1$  is given by

$$\text{var } \hat{\beta}_1 \approx \frac{1}{\sum_j \frac{1}{\hat{\sigma}_1^2 + V_{1j}}}.$$

Similarly the likelihood of obtaining  $\beta_{21}, \dots, \beta_{26}$  is

$$L = \prod_{j=1}^6 \frac{1}{\sqrt{2\pi} \sqrt{\sigma_2^2 + v_{2j}}} \exp \left[ -1/2 \frac{(\hat{\beta}_{2j} - \beta_2)^2}{(\sigma_2^2 + v_{2j})} \right],$$

and  $\hat{\sigma}_2^2$ ,  $\hat{\beta}_2$  as well as an estimate of the variance of  $\hat{\beta}_2$  can be obtained in the same way.

### Results

$$\hat{\beta}_1 = 0.0997, \hat{\beta}_2 = 0.0719.$$

$$\hat{\sigma}_1 = 0.0219, \hat{\sigma}_2 = 0.0089.$$

$$\text{Estimated SE}(\hat{\beta}_1) = 0.0098, \text{estimated SE}(\hat{\beta}_2) = 0.0043.$$

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{(\text{SE}(\hat{\beta}_1))^2 + (\text{SE}(\hat{\beta}_2))^2}} = 2.59$$

The same problems arise as for the case where we assumed  $\sigma_1 = \sigma_2$ , except that this time  $\hat{\beta}_1$  and  $\hat{\beta}_2$  aren't correlated. Again an exact p-value cannot be quoted, but again a significant difference between  $\beta_1$  and  $\beta_2$  is strongly suggested. Our test statistic will be found significant at the 5% level if it is compared to  $t_f$  tables for any  $f \geq 3$ . It seems to make very little difference whether or not we assume  $\sigma_1 = \sigma_2$ ; the result is almost identical for both cases. So the fact that for case 1,  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are correlated probably doesn't affect the analysis too much. However the fact remains that our analysis depends on our assumption of normality for the  $\beta_{1j}$ 's and may be sensitive to this assumption. A summary of the results is given below:

Estimate  $\pm$  standard error

Parameter	Naive Analysis	Second Analysis	
		Second Analysis assuming $\sigma_1 = \sigma_2$	without assuming $\sigma_1 = \sigma_2$
$\beta_1$	0.1019 $\pm$ .0041	0.0983 $\pm$ .0076	0.0997 $\pm$ .0098
$\beta_2$	0.0724 $\pm$ .0024	0.0722 $\pm$ .0068	0.0719 $\pm$ .0043

Our analyses suggest that  $\beta_1$  and  $\beta_2$  are significantly different which can be interpreted as meaning that system is an important factor in determining the survival distribution, at least of the oyster lines. Significant system effect is certainly not surprising as when the nonparametric tests were carried out the biggest differences occurred when lines from different systems were compared.

## 2.5 Incorporating Netting and Dowling Lines into the Model

It is not only the system effect which is of interest. It is also of interest to know whether the type of material plays a role in determining the survival distribution. With only one netting line in each system and one dowling line in system two it will not be possible to obtain an estimate of line-to-line variance for each type of material

within each system. It will have to be assumed that line-to-line variance is independent of type of material. As before  $\sigma_1^2$  and  $\sigma_2^2$  are the line-to-line variances for systems one and two respectively (regardless of material type).

The previous model is extended to incorporate netting and dowling lines as follows:

Let  $\beta_{ijk}$  be the true  $\beta$  value for the  $k$ th line of the  $j$ th type in system

$i$ . Let  $\hat{\beta}_{ijk}$  be our maximum likelihood estimate of  $\beta_{ijk}$  and let  $V_{ijk}$  = variance of  $\hat{\beta}_{ijk}$ , assumed to be a known constant. Assume

$$\hat{\beta}_{ijk} | \beta_{ijk} \sim N(\beta_{ijk}, V_{ijk})$$

and

$$\beta_{ijk} \sim N(\beta + f_i + t_j, \sigma_i^2)$$

where  $f_i$  = system effect  $i=1,2$ , and  $\sum_i f_i = 0$ ,

$t_j$  = type effect  $j=1,2,3$ , and  $\sum_j t_j = 0$ .

So  $t_1$  = effect due to type 1 (oyster),

$t_2$  = effect due to type 2 (dowling),

$t_3$  = effect due to type 3 (netting).

We are assuming no system-type interaction. The variance of  $\beta_{ijk}$  is  $\sigma_i^2$  instead of  $\sigma_{ij}^2$  as we don't have sufficient replication to allow the variance to depend on type. As before the assumption of normality for the  $\hat{\beta}_{ijk}$ 's is certainly reasonable, but the assumption of normality for the  $\beta_{ijk}$ 's is purely an assumption and cannot be verified. It follows that

$$\hat{\beta}_{ijk} \sim N(\beta + f_i + t_j, \sigma_i^2 + V_{ijk}).$$

In order to compare the slopes of the survival curves for the two systems and for the three types of material, we need estimates of  $\beta$ ,  $f_1$ ,

$f_2, t_1, t_2, t_3, \sigma_1^2$  and  $\sigma_2^2$  and estimates of the variances and covariances of  $f_1, f_2, t_1, t_2, t_3$ . The likelihood under our model of obtaining the seventeen  $\hat{\beta}_{ijk}$ 's that we obtained is

$$L = \prod_i \prod_j \prod_k \frac{1}{\sqrt{2\pi} \sqrt{\sigma_i^2 + V_{ijk}}} \exp \left[ -1/2 \frac{(\hat{\beta}_{ijk} - \beta - f_i - t_j)^2}{(\sigma_i^2 + V_{ijk})} \right]$$

Substituting  $f_2 = -f_1$  and  $t_3 = -t_1 - t_2$ , the maximum likelihood estimates  $(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{f}_1, \hat{t}_1, \hat{t}_2, \hat{\beta})$  are obtained by Newton Raphson. During the iteration,  $\sigma_1^2$ , and  $\sigma_2^2$  are constrained to be nonnegative.

### Results

$$\hat{\beta} = 0.0862.$$

$$\hat{f}_1 = 0.0176 \implies f_2 = -0.0176.$$

$$\hat{t}_1 = 0.00245, \hat{t}_2 = 0.00105, \hat{t}_3 = -0.00350.$$

$$\hat{\sigma}_1 = .0234, \hat{\sigma}_2 = .0072.$$

The  $\hat{\beta}_{ijk} = \hat{\beta} + \hat{f}_i + \hat{t}_j$  are given below:

System	Type	Oyster	Netting	Dowling
1		0.1063	0.1003	0.1049
2		0.0711	0.0651	0.0697

Note that the estimates of the slope for respectively the system one and

the system two oyster lines (0.1063 and 0.0711) are similar but not identical to the estimates obtained before under the model for the oyster lines alone (respectively 0.0997 and 0.0719).

Estimates of the variances of these estimates are obtained as usual using the matrix of second derivatives evaluated at the maximum likelihood estimates, which was already calculated during the Newton Raphson process.

$$\hat{SE} \hat{\beta} = .0035.$$

$$\hat{SE} \hat{f}_1 = \hat{SE} \hat{f}_2 = .0031.$$

$$\hat{SE} \hat{t}_1 = .0030, \hat{SE} \hat{t}_2 = .0038, \hat{SE} \hat{t}_3 = .0039.$$

It was then of interest to test for differences between  $f_1$  and  $f_2$  and between  $t_1$ ,  $t_2$  and  $t_3$ :

$$\frac{\hat{f}_1 - \hat{f}_2}{\sqrt{\hat{var}(\hat{f}_1 - \hat{f}_2)}} = 5.62$$

$$\frac{\hat{t}_1 - \hat{t}_2}{\sqrt{\hat{var}(\hat{t}_1 - \hat{t}_2)}} = 0.25$$

$$\frac{\hat{t}_1 - \hat{t}_3}{\sqrt{\hat{var}(\hat{t}_1 - \hat{t}_3)}} = 1.02$$

$$\frac{\hat{t}_1 - \hat{t}_2}{\sqrt{\hat{var}(\hat{t}_1 - \hat{t}_2)}} = 0.25$$

$$\frac{\hat{t}_1 - \hat{t}_3}{\sqrt{\hat{var}(\hat{t}_1 - \hat{t}_3)}} = 1.02$$

$$\frac{\hat{t}_1 - \hat{t}_2}{\sqrt{\hat{var}(\hat{t}_1 - \hat{t}_2)}} = 0.25$$

Since  $t_1 + t_2 + t_3 = 0$  it is unnecessary also to compare  $t_2$  to  $t_3$ . As before, it seems reasonable to compare these quantities to t-tables since the variances are estimated. Again it is not clear how many



degrees of freedom should be used. Since we are doing three comparisons the significance level should be adjusted accordingly: if the overall significance level is to be 5%, then significance level of each comparison could be  $5/3\% = 1.66\%$ . For the above comparisons, whatever the degrees of freedom we reach the same conclusion, namely that there is a significant difference between the  $\beta$  values for the two systems (as was found previously), but no significant difference between the  $\beta$  values for the three types of material. This suggests that regarding the rate at which the barnacles die off, there is a significant difference between the two systems but not between the three types of material.

### Discussion

Inspecting the  $\beta$  values, it is seen that the values for system one are generally larger than those for system two, whereas  $\beta$  values for any particular type of material show no tendency to be bigger or smaller than those for any other type. These observations are borne out by the above analyses which suggest a significant system effect but not a significant type effect. It is seen from  $\sigma_1$  and  $\sigma_2$  that  $\beta$  values are more variable in system one than in system two.

Statistical analyses relating to the  $\alpha$  values were not carried but an inspection of the values suggests that there is both a system and type effect. The  $\alpha$  values for system one are generally smaller than values for system two and values for dowling lines are generally smaller than values for the other types of line. The physical interpretation of

these observations is that barnacles on dowling and system one lines tend to start dying earlier than other barnacles. Barnacles in system one tend to die at a faster rate than those in system two. Furthermore the rate at which the system one barnacles die shows more variability from line to line.

The final observation relates to the initial number of barnacles on the lines. There are more barnacles initially on a line of a given type in system two than on a line of the same type in system one. Also we found the  $\hat{\beta}$  values to be generally smaller and the  $\hat{\alpha}$  values generally larger for system two lines than for system one lines. Formal statistical tests are not appropriate and even the simple correlation coefficient is not very meaningful with such a small number of lines of a given type. However it is worth noting that for the system one oyster lines, the correlation coefficient between the initial number of barnacles and  $\hat{\beta}$  is  $-.93$ . This correlation suggests that on lines with fewer barnacles initially, barnacles tend to start dying earlier and at a faster rate. However the correlation may also be due to a third variable related to both  $\hat{\beta}$  and initial number. This will be discussed again later in relation to the growth data. All in all it appears that system is more important than type as regards survival distribution. However this conclusion is somewhat tentative since it is based on the analysis of parameters  $\beta$  and  $\alpha$  of a model which doesn't adequately fit the data. Furthermore the subsequent analyses required a strong normality assumption.

### 3. EXAMINATION OF THE GROWTH DATA

#### 3.1 Growth Data

Growth data was collected at  $t = 1, 2, \dots, 9, 10, 11, 14, 17$  weeks. At each of these times samples were taken as follows: from each shell on each oyster line and from each piece of wood on each dowling line a haphazard sample of 5 live barnacles was selected. From each piece of rubber on each netting line a haphazard sample of 10 live barnacles was selected. It will be assumed that on each item every live barnacle was equally likely to be sampled, so for example the probability of being sampled was not related to size. For each sampled barnacle shell length and neck length were measured. Measurements were recorded in cm. The sampled barnacles were not removed from the items to which they were attached. In total there was an enormous amount of growth data: the total number of barnacles measured at each time was:

From oyster lines:	$2 \times 6 \times 10 \times 5 = 600,$
From dowling lines:	$3 \times 20 \times 5 = 300,$
From netting lines:	$2 \times 10 \times 10 = 200.$

The data was incomplete in several ways: at  $t = 1$  week, observations were taken only on barnacles from 3 oyster lines and 1 dowling line in system 2. Also at  $t = 10$  and 11 weeks observations were taken on barnacles from some of the lines but not all of them. It sometimes happened that fewer than five (or ten) barnacles were sampled from a particular item. This was because the population from which the

barnacles were sampled was continually getting smaller due to barnacles dying. Sometimes, towards the end of the study (particularly for the system one oyster lines where initial sample sizes were small) fewer than five (or ten) living barnacles remained on a particular item. In this case all remaining barnacles were sampled.

The analyses which were used assumed that measurements were taken at the same time points for all individuals. So only the data collected at  $t = 2, 3, 4, \dots, 9, 14, 17$  weeks was used. The analyses could have been extended to accommodate the missing data and methods of handling missing data will be discussed later. However this would have made the computation much more laborious and since the amount of data neglected is relatively small it wasn't considered worthwhile. Also items were not included if at any point in time fewer than five (or ten) barnacles were sampled from them. Again not much information was lost by doing this. Only the data relating to shell length was analyzed. The data relating to neck length could have been treated in an identical way. In summary: on each of 17 lines there are a number of items (oyster shells, pieces of wood or rubber). At each point in time, measurements are made on a random sample of barnacles taken from each item. The population from which these barnacles are sampled is continually diminishing because barnacles are dying.

The proportion of barnacles sampled on a particular item is approximately 10% initially and 20-30% towards the end of the study except for items on the system one oyster lines where the proportions are somewhat larger - approximately 25% initially and close to 100% by

the end of the study. These proportions are only estimates as we don't actually have the survival data for each item, only for each line.

Individual barnacles cannot be identified. We would like to fit growth curves for the shell length of the barnacles and to compare these curves for the two systems and for the three types of material.

### 3.2 Growth Curves for Items

Sophisticated growth curve models have been developed for longitudinal data - for the situation where an observation is obtained at each time point on each of a number of individuals. In our case we don't have longitudinal data for the barnacles: at each point we have measurements on a sample of unidentified barnacles - each barnacle may or may not have been measured at the previous time point. The items, however, can be identified and if we were to average over the sample obtained from each item at each time point, we could think of our data as longitudinal data for the items: on each line we would have a number of items and we would have a measurement representing each item at each time. This measurement would actually be the average shell length over a sample of barnacles picked from the remaining living barnacles on that item at that time.

Since each barnacle has equal probability of being sampled, regardless of size, and since the proportion of barnacles sampled is quite large, it should be reasonable to allow this average to represent all the remaining living barnacles on that items at that time. So

instead of following individual barnacles we would be following individual items.

Suppose we would like to fit polynomial growth curves. With effectively only one observation per item we will have to assume that all items within a line have the same parameters. So comparison of items will be impossible. But in any case the main interest lies in comparison of the 17 lines. We may allow each of the 17 lines to have different parameters and we may think of the items on each line as a sample from a conceptual population of items on that line. We may then fit the average growth curve for items on each line. What will this curve represent?

Consider the growth curve for the first oyster line. This will show how the average shell length of the remaining living barnacles on a typical item from this line changes with time. (We hope that the average over all barnacles on a particular item at a particular time will be well-represented by the average over a sample of barnacles taken from that item at that time). This average size changes for two reasons - firstly because the barnacles grow and secondly because the population of living barnacles changes due to barnacles dying. So the growth curve represents the natural population rather than individual barnacles. We would expect the growth curves for the natural population to be similar to the growth curves for individual barnacles, provided that the probability of dying is unrelated to size. If the barnacles that die in a particular time-interval are neither particularly big nor particularly small, then the change in population will not affect the average size of

the living barnacles. The changes in this average size will then be entirely due to the growth of the barnacles. If, on the other hand, there is a tendency for the larger barnacles, say, to die first, then while each individual barnacle grows bigger, the average size of the living barnacles might actually get smaller. If this were the case, the growth characteristics of the natural population would not reflect the growth characteristics of individual barnacles. Without being able to identify individual barnacles it is difficult to ascertain whether the probability of dying is related to size. This question will be addressed later along with a discussion of the legitimacy of the approach and of the assumptions that are implicitly being made.

So using the "longitudinal" data for items, polynomial growth curves were fitted for each of the 17 lines. Two growth curve models were employed - the first was due to Pothoff and Roy (1964) and the second to Rao (1965, 1966). The curves obtained for the two different systems and for the three types of material were then compared.

### **3.3 The Growth Curve Models**

#### **Pothoff and Roy's Model**

It is assumed that longitudinal data is available - successive measurements on a number of individuals. The measurements for any one individual are then clearly correlated. Typically the individuals are divided into groups which are to be compared.

It is assumed that each individual has measurements taken at the

same  $q$  points in time. The set of  $q$  measurements for any one individual form one row of the data matrix  $Y_0$ . The model is as follows:

$$E[Y_0] = \begin{matrix} & A & \xi & P \\ \begin{matrix} nxq & nxm & mxp & pxq \end{matrix} \end{matrix}$$

where  $n$  = number of individuals,

$m$  = number of groups,

$q$  = number of time points,

$p$  = number of parameters fitted within each group ( $p \leq q$ ),

and  $A$  is the design matrix across individuals,

$\xi$  is a matrix of parameters to be estimated,

$P$  is a matrix of known constants related to time (assumed to be of full rank,  $p$ ).

It is assumed that the rows of  $Y_0$  are mutually independent (measurements on different individuals are independent) and that the  $q$  elements in any one row follow the multivariate normal distribution with unknown covariance matrix  $\Sigma_0$  ( $q \times q$ ), the same for every individual.  $\Sigma_0$  is not diagonal since successive measurements on the same individual are correlated. It would be possible to extend the model to allow the covariance matrices for different individuals to be multiples of each other and also to allow different individuals to have measurements taken at different times. These extensions are appropriate in our situation and will be considered later.

To illustrate the model consider the following situation: we have  $m$  groups of individuals with  $n_j$  in the  $j$ th group. The growth curves for individuals in all groups are from the same family, for example



polynomials of degree  $p-1$ . But the parameters may be different for the different groups. So the growth curve associated with the  $j$ th group is

$$\xi_{j0} + \xi_{j1} t + \xi_{j2} t^2 + \dots + \xi_{j \ p-1} t^{p-1}$$

$$\text{Then } A = \begin{matrix} n_1 \{ \\ n_2 \{ \\ n_m \{ \end{matrix} \begin{bmatrix} 1 & 0 & . & . & . & . & 0 \\ 1 & 0 & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 1 & . & . & . & . & 0 \\ 0 & 1 & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & 0 & 1 \\ 0 & . & . & . & . & . & 0 & 1 \end{bmatrix}$$

$$\xi = \begin{bmatrix} \xi_{10} & . & . & . & . & \xi_{1 \ p-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \xi_{m0} & . & . & . & . & \xi_{m \ p-1} \end{bmatrix}$$

$$P = \begin{bmatrix} t_1^0 & . & . & . & . & t_q^0 \\ t_1^1 & . & . & . & . & t_q^1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ t_1^{p-1} & . & . & . & . & t_q^{p-1} \end{bmatrix}$$

Alternatively if the time points are equally spaced, orthogonal coefficients may be used for  $P$ .

This can be generalised to the situation where there are two or more factors each with a number of levels or to the multi-response

situation where more than one growth characteristic is measured. Any of the usual factorial or block designs can be accommodated by this model.

We want to estimate the parameters  $\xi_{ij}$  and to test various hypotheses of the form

$$\begin{matrix} C & \xi & V & = & 0. \\ s \times m & m \times p & p \times u \end{matrix}$$

For example, if the hypothesis is that all  $m$  growth curves are the same, i.e. that all parameters for all groups are equal, then

$$C = \begin{pmatrix} I & -1 \\ (m-1) \times (m-1) & (m-1) \times 1 \end{pmatrix} \text{ and } V = \begin{pmatrix} I \\ p \times p \end{pmatrix}.$$

If the hypothesis is that all curves are of degree  $p-2$  or less then

$$C = \begin{matrix} I \\ m \times m \end{matrix} \text{ and } V = \begin{matrix} (p \times 1) \\ \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{vmatrix} \end{matrix}$$

### Solution

Pothoff and Roy suggest a transformation which reduces the model to a more standard model which has been treated extensively in the literature. The transformation is

$$\begin{matrix} X & = & Y_0 & G^{-1} & P' & (P G^{-1} P')^{-1} \\ n \times p & & n \times q & q \times q & q \times p & p \times p \end{matrix}$$

where  $G$  is an arbitrary symmetric positive definite matrix such that  $P G^{-1} P'$  is of full rank. Then  $X$  is such that:

- (a) its rows are mutually independent,
- (b) the  $p$  elements in any one row follow the multivariate normal distribution with unknown positive definite covariance matrix

$$\Sigma_{p \times p} = [P(G')^{-1} P']^{-1} P(G')^{-1} \Sigma_0 G^{-1} P' (P G^{-1} P')^{-1},$$

(c)  $E[X] = A\xi$ .

This is the usual multivariate analysis of variance (MANOVA) model. Under this model a test of  $C\xi V = 0$  is based on  $S_h(u \times u)$  and  $S_e(u \times u)$  where

$$\begin{aligned} S_h &= (C \hat{\xi} V)' [C_1(A_1' A_1)^{-1} C_1']^{-1} (C \hat{\xi} V) \\ &= V' X' A_1(A_1' A_1)^{-1} C_1 [C_1(A_1' A_1)^{-1} C_1']^{-1} C_1(A_1' A_1)^{-1} A_1' X \end{aligned}$$

and 
$$S_e = V' X' [I - A_1(A_1' A_1)^{-1} A_1'] X V,$$

where  $\hat{\xi}$  is the least squares estimate of  $\xi$ ,

$A_1$  = first  $r$  columns of  $A$ ,

$C_1$  = first  $r$  columns of  $C$ ,

and  $\text{rank } A = r$  ( $\leq m, < n$ ),

$\text{rank } C = s$  ( $\leq m, \leq r$ ),

$\text{rank } V = u$  ( $\leq p$ ).

Several possible tests are available:

(1) Roy's test is based on the largest characteristic root of  $S_h S_e^{-1}$ ,

(2) Hotelling's test is based on the trace of  $S_h S_e^{-1}$ ,

(3) Wilk's test is based on the ratio of determinants  $\frac{|S_e|}{|S_h + S_e|}$ .

In order to test a hypothesis  $C\xi V = 0$  under the original model,  $X = Y_0 G^{-1} P' (P G^{-1} P')^{-1}$  is substituted into the expressions for  $S_h$  and  $S_e$ . In all analyses which follow, Roy's test was used. This proceeds as follows:

Let  $\lambda =$  largest characteristic root of  $S_h S_e^{-1}$ . Compare  $\frac{\lambda}{1 + \lambda}$  to Heck tables (e.g. Morrison, 1976) with parameters

$$\begin{aligned} s^* &= \min (s, u), \\ m^* &= 1/2 (|s-u|-1), \\ n^* &= 1/2 (n-r-u-1). \end{aligned}$$

Reject the null hypothesis if  $\frac{\lambda}{1 + \lambda}$  is too large.

If  $s^* = 1$ , the statistic  $\frac{(n^* + 1)}{(m^* + 1)} \lambda$  can be compared to F tables with

$(2m^* + 2)$  and  $(2n^* + 2)$  degrees of freedom. The least squares estimator of  $\xi$  is given by

$$\hat{\xi} = (A_1' A_1)^{-1} A_1' X = (A_1' A_1)^{-1} A_1' Y_0 G^{-1} P' (P G^{-1} P')^{-1}.$$

The best linear unbiased estimator of the function  $C\xi V$  is then  $C\hat{\xi}V$ .

Confidence bounds can also be found for these functions.

### Choice of G

The analysis is valid for any G satisfying the given conditions. However the choice of G affects the power of the tests and the variance of the estimators. The minimum variance unbiased estimator of  $\xi$  is  $(A_1' A_1)^{-1} A_1' Y_0 \Sigma_0^{-1} P' (P \Sigma_0^{-1} P')^{-1}$ . Comparing this with the least squares estimator of  $\xi$  given above, would suggest that the optimal choice of G is  $G = \Sigma_0$ . Pothoff and Roy suggest that the more G differs from  $\Sigma_0$  the worse the power of the test will be and the greater the variance of the estimators. However  $\Sigma_0$  is unknown and an estimate of  $\Sigma_0$  obtained from  $Y_0$  may not be used. If the

experimenter has no idea about the form of  $\Sigma_0$  the simplest procedure is to use  $G=I$ . If an estimate of  $\Sigma_0$  can be obtained from a similar but independent experiment it would be preferable to use this estimate. Alternatively, if a guess of  $\Sigma_0$  can be made before the experiment is run, this may be used. For example, it might be assumed that the correlation between any two observations  $d$  periods apart is  $\rho^d$  and that the variance is constant with respect to time. Then  $\Sigma_0$  is proportional to

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \dots \\ \rho & 1 & \rho & \rho^2 \dots \\ \rho^2 & \rho & 1 & \rho \dots \\ \rho^3 & \rho^2 & \rho & 1 \dots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Unfortunately, however, the choice of  $\rho$  is arbitrary unless an estimate of it can be obtained from an independent experiment.

### Rao's Model

The problem with Pothoff and Roy's model is that  $G$  must be chosen arbitrarily. Rao points out that their model doesn't utilize all the available information unless  $G = \Sigma_0$  and they do not allow an estimate of  $\Sigma_0$  obtained from the data to be used. So Rao suggests an alternative way of reducing the growth model to the usual MANOVA model: construct a  $q \times q$  nonsingular matrix  $H = (H_1 | H_2)$  such that the columns

of  $H_1$  form a basis for the vector space spanned by the rows of  $P$  and  $PH_1 = I$ ,  $PH_2 = 0$ . When  $\text{rank } P = p$  we can choose

$$H_1 = G^{-1} P' (P G^{-1} P')^{-1}, H_2 = I - H_1 P$$

where  $G$  is an arbitrary positive definite matrix. The choice of  $G$  does not affect estimates or tests.

$$\text{Let } \underset{n \times p}{X} = \underset{n \times o}{Y_o} \underset{o \times p}{H_1} = \underset{n \times o}{Y_o} \underset{o \times p}{G^{-1} P' (P G^{-1} P')^{-1}} \text{ and } \underset{n \times (q-p)}{Z} = \underset{n \times o}{Y_o} \underset{o \times (q-p)}{H_2}.$$

Then, as for the Pothoff and Roy's model,  $E[X] = A\xi$ , but in addition  $E[Z] = 0$ . Hence  $E[X|Z] = A\xi + Z\Gamma$  where  $Z$  is a matrix of  $q-p$  covariables and  $\underset{(q-p) \times p}{\Gamma}$  is a matrix of unknown regression coefficients.

Rao claims that the estimate of  $\xi$  obtained under this model is more efficient than that obtained under Pothoff and Roy's model as it uses information contained in the covariables  $Z$ , which is neglected under Pothoff and Roy's model. However this is not true. The best linear unbiased estimator of  $\xi$  under Rao's model is

$$\hat{\xi} = (A_1' A_1)^{-1} A_1' Y_o S^{-1} P' (P S^{-1} P')^{-1}$$

where  $S$  is an estimate of  $\Sigma_o$  obtained from the data:

$$S = Y_o' [I - A_1 (A_1' A_1)^{-1} A_1'] Y_o$$

This is precisely what we would obtain under Pothoff and Roy's model if we set  $G = S$ , thus allowing an estimate of  $\Sigma_o$  obtained from the data as Rao does.

To test hypotheses of the form  $C\xi V = 0$ , matrices  $S_h$  and  $S_e$  are again found. This time they have a slightly different form:

$$\text{Let } B = [S^{-1} - S^{-1} P' (P S^{-1} P')^{-1} P S^{-1}].$$

$$\text{Then } S_h = V' X' A_1 (A_1' A_1)^{-1} C_1' (C_1 R C_1')^{-1} C_1 (A_1' A_1)^{-1} A_1' X V$$

$$S_e = V' (P S^{-1} P')^{-1} V$$

$$\text{where } R = (A_1' A_1)^{-1} + (A_1' A_1)^{-1} A_1' Y_0' B Y_0' A_1 (A_1' A_1)^{-1}$$

$$\text{and } X = Y_0 S^{-1} P' (P S^{-1} P')^{-1}.$$

The expression for  $S_e$  is identical to the expression for  $S_e$  in Pothoff and Roy's model with  $G$  set equal to  $S$ . However  $S_h$  is not the same. The tests based on  $S_h$  and  $S_e$  are the same as for Pothoff and Roy's model. However when comparing  $\frac{\lambda}{1 + \lambda}$  to Heck tables the parameter  $n^*$  is now slightly different,

$$n^* = n - s - u - 1 - (q-p).$$

Sometimes as discussed by Grizzle and Allen (1969) it is preferable to use fewer than  $q-p$  covariables.

### 3.4 Application of the Models

The two growth curve models were applied to the longitudinal data for items. Each row of the data matrix contained nine entries. The  $i$ th entry was the average shell length over a random sample of living barnacles taken at  $t = t_1$  from the item corresponding to that row. The items were divided into groups according to the line that they were on. In their present form the models require the following assumptions:

- (1) The rows of the data matrix are mutually independent.
- (2) The set of entries in any row follows the multivariate normal distribution with the same covariance matrix  $\Sigma_0$  for each row.

Provided that we are prepared to assume that observations on barnacles from different items are independent, then the rows of our data matrix will certainly be mutually independent since each row of our data matrix corresponds to a different item and no two items have any barnacles in common. Furthermore, if we are prepared to assume that the set of observations on an individual barnacle is multivariate normal then the first part of assumption 2 would also be satisfied since each entry in our data matrix would then be the average of a number of variables assumed to be normally distributed. Whether or not it is reasonable to assume that the covariance matrix is the same for every row of our data matrix will be discussed later.

Subsequently another analysis was carried out in which the weaker and more reasonable assumption was made that a set of observations corresponding to items from oyster or dowling lines had covariance matrix  $\frac{1}{5} \Sigma$  and a set of observations corresponding to items from netting lines had covariance matrix  $\frac{1}{10} \Sigma$  for some fixed, unknown  $\Sigma$ . In order to do this the model had to be modified slightly as described in Section 3.8. It is not clear that even this weaker assumption is reasonable. This too will be discussed later. Only those items for which every observation was an average over five (ten) barnacles were included in the analysis.

In total 141 items were used in the analyses. The number of items used from each line was as follows:



01	5	D1	11	010	6
02	8	D2	11	011	10
03	5	N1	8	012	9
04	6	07	9	D3	13
05	6	08	8	N2	10
06	8	09	8		

The main objective was to fit polynomial growth curves to each of the lines and to determine whether differences existed between these curves for the three different types of materials and for the two systems. As a first step it was of interest to determine whether lines within each system-type combination could be ignored. If this were the case it would be reasonable to collapse over lines of the same type within the same system and we would have a two factor situation with the following number of items.

Type	System 1	System 2
Oyster	38	50
Dowling	22	13
Netting	8	10

In order to determine whether differences existed between the growth curves for the six oyster lines in system one, polynomial growth curves were fitted to the six lines and the parameters were compared. A similar analysis was carried out for the six oyster lines in system two.



where  $p$  is the number parameters fitted. Since we have taken the  $(j,k)$ th element of  $P$  to be  $(t_k - \bar{t})^{j-1}$  rather than  $t_k^{j-1}$ , the growth curve for the  $j$ th group will be  $\xi_{0j} + \xi_{1j} (t - \bar{t})$   
 $+ \dots + \xi_{p-1j} (t - \bar{t})^{p-1}$ .

The items are divided into six groups with the following number in each group: 5, 8, 5, 6, 6, 8. So

$$A = \begin{matrix} 5\{ \\ (38 \times 6) \\ 8\{ \\ 8\{ \end{matrix} \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & 0 & 1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & 0 & 1 \end{bmatrix}$$

$n = 38$  (number of individuals),  $m = 6$  (number of groups),  $q = 9$ ,  $r = 6$ .

In Section 3.3, it was described how to test hypotheses of the form  $C\xi V = 0$  under each model. By choosing  $C$  and  $V$  appropriately the following analyses were carried out under each model:

1. It was determined what degree of polynomial adequately fitted the data. This was done in the following way: initially cubics were fitted and the hypothesis that all curves were of degree 2 or less was tested.

This required a test of  $\xi_{13} = \xi_{23} = \dots = \xi_{63} = 0$  which can be written  $C\xi V = 0$  with  $C = I$  (6 x 6) and

$$V = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, s = 6, u = 1.$$

If quadratics were found to be adequate, quadratics were then fitted in order to determine whether linear curves would also be adequate. If quadratics were not found to be adequate, quartics were then fitted to determine whether cubics would be adequate. This process continued until the appropriate polynomial was determined.

2. The hypothesis that all six growth curves were identical was tested. This hypothesis can be written

$$\begin{aligned} \xi_{10} &= \xi_{20} = \dots = \xi_{60} \\ \xi_{11} &= \xi_{21} = \dots = \xi_{61} \\ &\cdot \\ &\cdot \\ &\cdot \\ \xi_{1 \ p-1} &= \xi_{2 \ p-1} = \dots = \xi_{6 \ p-1} \end{aligned}$$

or  $C\xi V = 0$  where  $C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$   
(5x6)

and  $V = I (p \times p)$ ,  $s = 5$ ,  $u = p$ ,

where  $p-1$  is the degree of the polynomial found to be appropriate in 1.

3. The hypothesis that the curves were identical except possibly for an additive constant (i.e. parallel) was tested. This hypothesis can be written  $C\xi V = 0$  where  $C$  is as in 2. and

$$V = \left| \begin{array}{c} 0 \\ \hline I \\ (p-1) \times (p-1) \end{array} \right|, s = 5, u = p-1.$$

### Results

1. (i) Cubics were fitted and the hypothesis that all curves were of degree 2 or less was tested.

a. Pothoff and Roy's model with  $G = \hat{\Sigma}_0$

$$S_h S_e^{-1} = .115, \text{ refer } 1.04 \text{ to } F_{6,54}, p > .1$$

b. Rao's model

$$S_h S_e^{-1} = .138, \text{ refer } 1.12 \text{ to } F_{6,49}, p > .1$$

According to both models, quadratics are adequate. It is possible that linear curves are also adequate.

(ii) Quadratics were fitted and the hypothesis that all curves were linear was tested.

a. Pothoff and Roy's with  $G = I$

$$S_h S_e^{-1} = 1.84, \text{ refer } 16.6 \text{ to } F_{6,54}, p < .001$$

b. Pothoff and Roy's model with  $G = \hat{\Sigma}_O$

$$S_h S_e^{-1} = 1.41, \text{ refer 12.7 to } F_{6,54}, p < .001$$

c. Rao's model

$$S_h S_e^{-1} = 1.86, \text{ refer 14.3 to } F_{6,46}, p < .001$$

Clearly linear curves are not adequate and quadratics are appropriate. So for each line we have estimates of three parameters, which are as follows:

a. Pothoff and Roy,  $G = I$

$$\hat{\xi} = \begin{vmatrix} 2.00 & .113 & - .0043 \\ 1.95 & .119 & - .0050 \\ 1.93 & .127 & - .0063 \\ 1.99 & .117 & - .0054 \\ 2.05 & .106 & - .0069 \\ 2.00 & .112 & - .0049 \end{vmatrix}$$

b. Pothoff and Roy,  $G = \hat{\Sigma}_O$

$$\hat{\xi} = \begin{vmatrix} 2.00 & .114 & - .0042 \\ 1.95 & .116 & - .0047 \\ 1.91 & .123 & - .0057 \\ 1.98 & .120 & - .0051 \\ 2.09 & .102 & - .0066 \\ 2.01 & .107 & - .0044 \end{vmatrix}$$

c. Rao

$$\hat{\xi} = \begin{vmatrix} 1.97 & .112 & - .0043 \\ 1.92 & .119 & - .0046 \\ 1.94 & .132 & - .0074 \\ 2.02 & .126 & - .0070 \\ 2.02 & .104 & - .0060 \\ 1.99 & .114 & - .0045 \end{vmatrix}$$

All three models give very similar estimates of  $\xi$ . In particular, the choice of  $G$  in Pothoff and Roy's model affects the estimate very little. So, letting  $y(t)$  denote shell length at time  $t$ , the growth curves that we obtain under, say, Rao's model for 01, 02, ..., 06 are respectively:

$$y(t) = - .0043 t^2 + .167 t + 1.07$$

$$y(t) = - .0046 t^2 + .178 t + .96$$

$$y(t) = - .0074 t^2 + .227 t + .78$$

$$y(t) = - .0070 t^2 + .216 t + .92$$

$$y(t) = - .0060 t^2 + .181 t + 1.10$$

$$y(t) = - .0045 t^2 + .172 t + 1.07$$

According to Rao's model, estimates of the average shell length over a sample of living barnacles taken from an item on 01 at  $t = 2, 3, 4, 5, 6, 7, 8, 9, 14$  weeks are respectively: 1.39, 1.53, 1.67, 1.80, 1.92, 2.03, 2.13, 2.23, 2.57 cm. Pictures of the six curves obtained under Rao's model are given in Figure 1. The six curves are not too far from being linear and are very similar to each other - almost parallel, although not quite identical. The hypotheses that the curves were identical and that they were identical except for an additive constant were then tested with the following results:

2. The hypothesis that all six curves were identical was tested.

a. Pothoff and Roy's model,  $G = I$

Figure 1. Growth Curves for the System 1 Oyster Lines

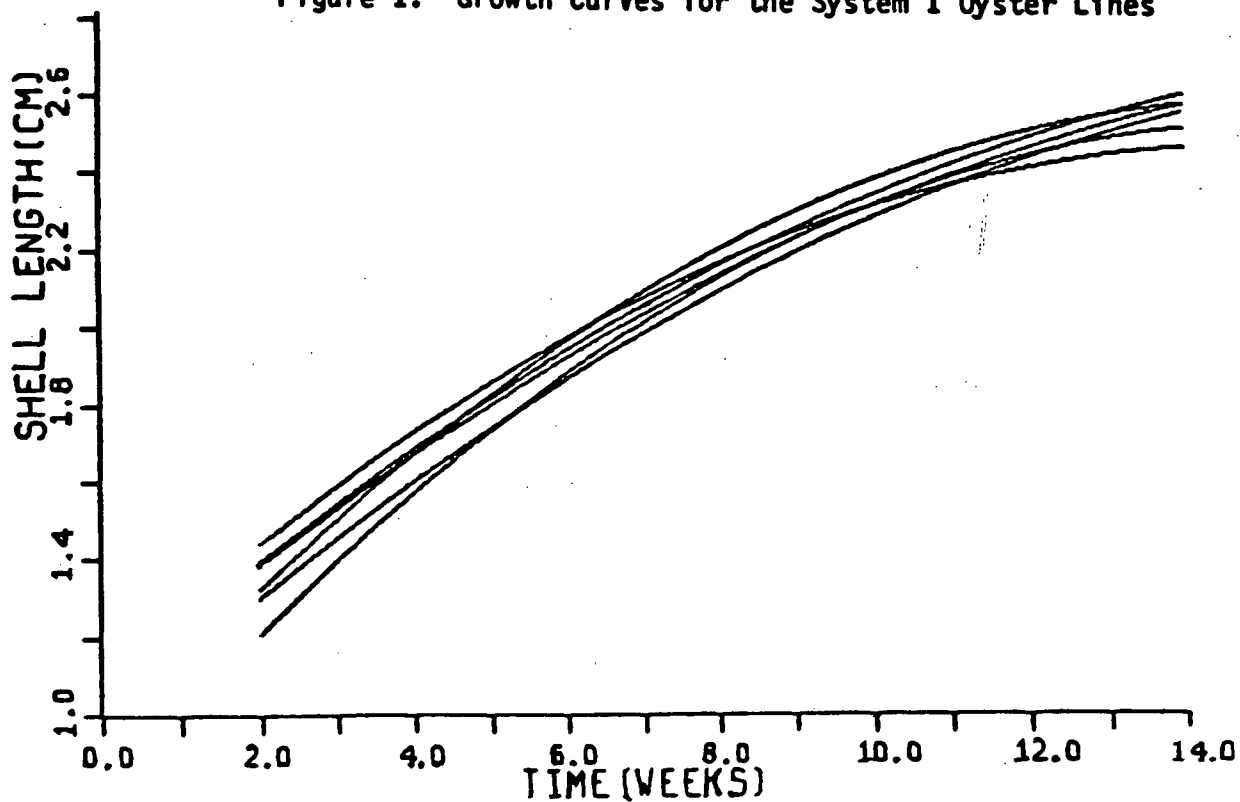
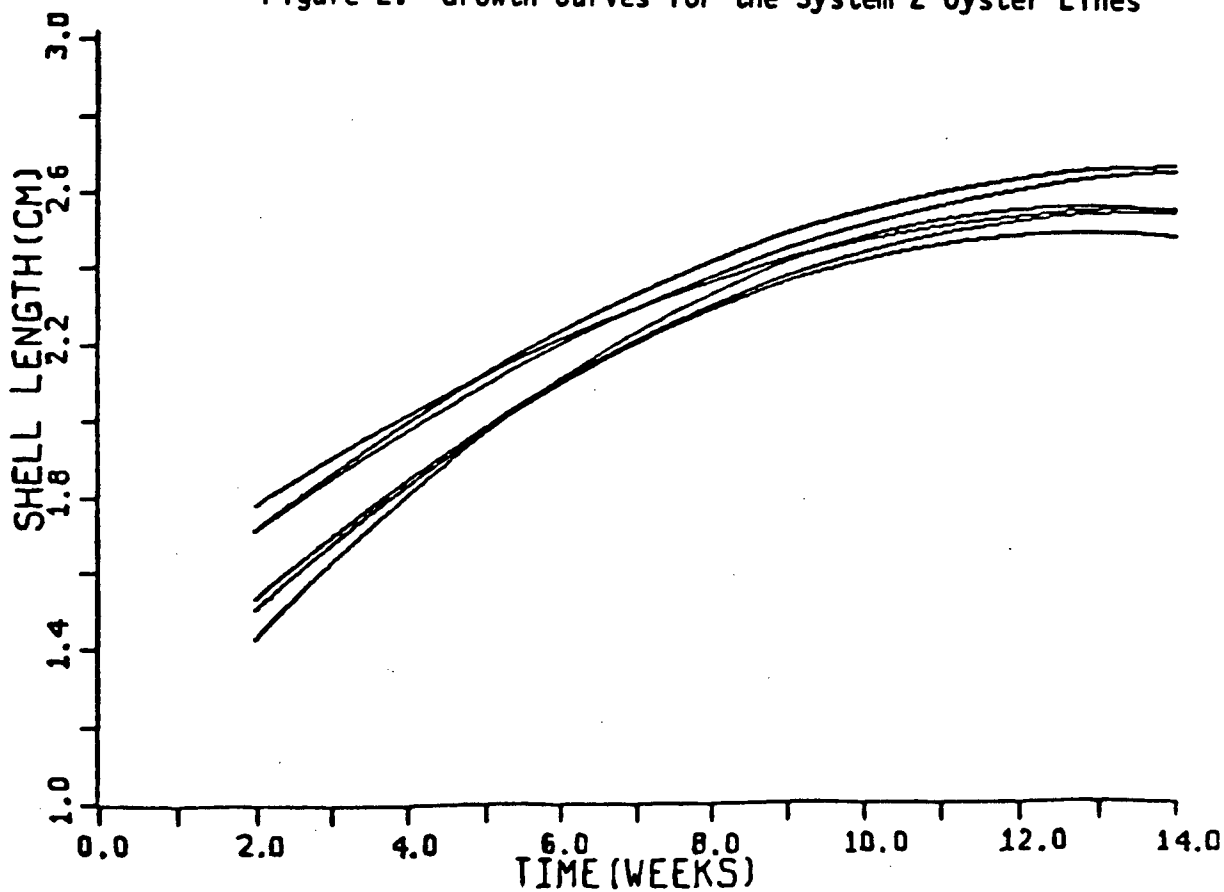


Figure 2. Growth Curves for the System 2 Oyster Lines





Largest root of  $S_h S_e^{-1} = .419$

Refer  $\frac{.419}{1.419} = .29$  to Heck tables

with  $s^* = 3$ ,  $m^* = 1/2$ ,  $n^* = 25$ .

$p \approx .04$

b. Pothoff and Roy's model,  $G = \hat{\Sigma}_0$

Largest root of  $S_h S_e^{-1} = .613$

Refer  $\frac{.613}{1.613} = .38$  to Heck tables

with  $s^* = 3$ ,  $m^* = 1/2$ ,  $n^* = 25$ .

$p \approx .01$

c. Rao's model

Largest root of  $S_h S_e^{-1} = .452$

Refer  $\frac{.452}{1.452} = .31$  to Heck tables

with  $s^* = 3$ ,  $m^* = 1/2$ ,  $n^* = 25$ .

$p \approx .05$

We have a sufficient amount of data to reject the hypothesis that the six growth curves are identical, although they appear very similar in Figure 1. There is a considerable discrepancy between the p-values obtained under Pothoff and Roy's model with  $G = I$ , and with  $G = \hat{\Sigma}_0$ . It seems that the choice of  $G$  greatly affects the variance of the estimators and hence the p-values.

3. The hypothesis that the curves were identical except for an additive constant was tested.

a. Pothoff and Roy's model,  $G = I$

Largest root of  $S_h S_e^{-1} = .164$

Refer  $\frac{.164}{1.164} = .14$  to Heck tables

$p \gg .05$

b. Pothoff and Roy's model,  $G = \hat{\Sigma}_0$

Largest root of  $S_h S_e^{-1} = .224$

Refer  $\frac{.224}{1.224} = .18$  to Heck tables

$p \gg .05$

c. Rao's model

Largest root of  $S_h S_e^{-1} = .212$

Refer  $\frac{.212}{1.212} = .17$  to Heck tables

$p \gg .05$

Not surprisingly the curves are found to be parallel. This conclusion is clear in spite of the considerable discrepancy which is again observed between the p-values obtained under Pothoff and Roy's model with  $G = \hat{\Sigma}_0$  and with  $G = I$ . Before discussing the results further, similar results will now be obtained for the system two oyster lines.

### 3.6 System Two Oyster Lines

The same analyses were repeated for the system two oyster lines.

We had a new data matrix  $Y(50 \times 9)$  and a new design matrix

$$A = 9 \left\{ \begin{array}{cccccc} 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & 1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & 1 \end{array} \right\}$$

while  $P$  was unchanged.

Again the analyses were carried out using Rao's model and using Pothoff and Roy's model both with  $G = I$  and with  $G =$  an estimate of  $\Sigma_0$  obtained from the system one oyster lines,  $\hat{\Sigma}_0 = Y_0' [I - A_1 (A_1' A_1)^{-1} A_1'] Y_0$  where  $Y_0$  and  $A$  are respectively the data matrix and the design matrix for the system one oyster lines.

### Results

1. (i) Cubics were fitted and the hypothesis that all curves were of degree 2 or less was tested.

- a. Pothoff and Roy's model with  $G = \hat{\Sigma}_0$

$$S_h S_e^{-1} = .066, \text{ refer } .59 \text{ to } F_{6,54}, p > .1$$

- b. Rao's model

$$S_h S_e^{-1} = .196, \text{ refer } 1.6 \text{ to } F_{6,49}, p > .1$$

Clearly quadratics are adequate. The conclusion is clear despite the discrepancy between the two values of the test statistic.

(ii) Quadratics were then fitted and the hypothesis that all curves were linear was tested.

a. Pothoff and Roy,  $G = I$

$$S_h S_e^{-1} = 8.37, \text{ refer } 75.3 \text{ to } F_{6,54}, p \ll .001$$

b. Pothoff and Roy  $G = \hat{\Sigma}_O$

$$S_h S_e^{-1} = 4.51, \text{ refer } 40.5 \text{ to } F_{6,54}, p \ll .001$$

c. Rao's model

$$S_h S_e^{-1} = 8.26, \text{ refer } 62.9 \text{ to } F_{6,46}, p \ll .001$$

The hypothesis that linear curves are adequate is very strongly rejected. So quadratics are again appropriate. The estimates of  $\xi$  are:

a. Pothoff and Roy,  $G = I$

$$\hat{\xi} = \begin{vmatrix} 2.23 & .094 & - .0052 \\ 2.15 & .104 & - .0070 \\ 2.28 & .076 & - .0053 \\ 2.17 & .125 & - .0099 \\ 2.28 & .097 & - .0063 \\ 2.16 & .106 & - .0082 \end{vmatrix}$$

b. Pothoff and Roy,  $G = \hat{\Sigma}_O$

$$\hat{\xi} = \begin{vmatrix} 2.21 & .091 & - .0042 \\ 2.16 & .101 & - .0069 \\ 2.29 & .077 & - .0059 \\ 2.18 & .128 & - .0105 \\ 2.27 & .095 & - .0063 \\ 2.14 & .103 & - .0080 \end{vmatrix}$$

c. Rao

$$\hat{\xi} = \begin{vmatrix} 2.24 & .094 & - .0054 \\ 2.16 & .107 & - .0074 \\ 2.25 & .081 & - .0055 \\ 2.17 & .122 & - .0096 \\ 2.28 & .099 & - .0064 \\ 2.15 & .107 & - .0084 \end{vmatrix}$$

Again, the estimates obtained from all three methods are almost identical. The growth curves that we obtain under Rao's model for 07, 08, ..., 012 are respectively:

$$y(t) = - .0054 t^2 + .164 t + 1.41$$

$$y(t) = - .0074 t^2 + .202 t + 1.16$$

$$y(t) = - .0055 t^2 + .152 t + 1.50$$

$$y(t) = - .0096 t^2 + .246 t + 0.98$$

$$y(t) = - .0064 t^2 + .181 t + 1.38$$

$$y(t) = - .0084 t^2 + .215 t + 1.11$$

According to Rao's model, estimates of the average shell length over a sample of living barnacles taken from an item on 07 at  $t = 2, 3, 4, 5, 6, 7, 8, 9, 14$  weeks are respectively: 1.72, 1.85, 1.98, 2.10, 2.20, 2.29, 2.38, 2.45, 2.65 cm. Pictures of the six curves obtained under Rao's model are given in Figure 2. The curves are less linear than those obtained for the system one oyster lines (the hypothesis of linearity was rejected with a much smaller p-value for the system two oyster lines than for the system one oyster lines). They are also less homogeneous than the curves obtained for the system one oyster lines -

whereas the curves for the system one oyster lines were found to be parallel, these curves are far from being parallel. This observation was confirmed by the results of the following tests:

2. The hypothesis that all six curves were identical was tested.

a. Pothoff and Roy's model,  $G = I$

Largest root of  $S_h S_e^{-1} = .801$

Refer  $\frac{.801}{1.801} = .44$  to Heck tables

1% critical value = .32,  $p < .01$

b. Pothoff and Roy's model,  $G = \hat{\Sigma}_O$

Largest root of  $S_h S_e^{-1} = .894$

Refer  $\frac{.894}{1.894} = .47$  to Heck tables, 1% critical value = .32,

1% critical value = .32,  $p < .01$

c. Rao's model

Largest root of  $S_h S_e^{-1} = .660$

Refer  $\frac{.660}{1.660} = .40$  to Heck tables

1% critical value = .36,  $p < .01$

As expected the curves are not found to be identical.

3. The hypothesis that the curves were parallel was then tested.

a. Pothoff and Roy's model,  $G = I$

Largest root of  $S_h S_e^{-1} = .648$

Refer  $\frac{.648}{1.648} = .39$  to Heck tables

1% critical value = .30,  $p < .01$

b. Pothoff and Roy's model,  $G = \hat{\Sigma}_0$

Largest root of  $S_h S_e^{-1} = .678$

Refer  $\frac{.678}{1.678} = .40$  to Heck tables,

1% critical value = .30,  $p < .01$

c. Rao's model

Largest root of  $S_h S_e^{-1} = .490$

Refer  $\frac{.49}{1.49} = .33$  to Heck tables,

1% critical value = .33,  $p \approx .01$

The p-values are somewhat larger than in 2., but nevertheless, even the hypothesis that the curves are parallel must be rejected.

### Discussion

As already observed, the growth curves for the system one oyster lines are more homogenous than the curves for the system two oyster lines. But even for the system one lines differences were found, so it won't be reasonable to ignore lines within a particular system-type combination. The constant term is generally larger for the system two curves than for the system one curves which suggests that barnacles on system two are generally larger than those on system one. The system

one curves are more linear than the system two curves. The average size of barnacles on items in system two appears to increase fairly quickly to start with and much less rapidly later on, whereas the average size of barnacles on items in system one increases at a more constant rate. We cannot conclude however, that barnacles on system two tend to grow more quickly at first than barnacles on system one and less quickly later on, as the curves represent the natural population, not individual barnacles. It is possible that the difference does reflect different growth characteristics for the two systems, but if the probability of dying is related to size, it may also reflect the different survival patterns for the two systems. Suppose, for example, that there is a tendency for the bigger barnacles to die first, then a large number of deaths in a particular interval would cause the increase in average size in that interval to be smaller than would be expected for the increase in size of any individual. On system one oyster lines barnacles die off very quickly between  $t = 5$  and  $t = 8$  weeks, but at a much slower rate after that, whereas barnacles on system two oyster lines die off at a more constant rate between  $t = 5$  and  $t = 17$  weeks. This could account for the more linear curves for system one lines.

### The Models

The estimates of  $\xi$  are very similar for all three methods. As already pointed out, the choice of  $G$  in Pothoff and Roy's model doesn't greatly affect the estimate of  $\xi$ , but it does affect its variance and hence the p-values of the tests. The p-value obtained using  $G = I$



was usually considerably different from that obtained using  $G = \hat{\Sigma}_0$ .

The p-value obtained with  $G = I$  tended to agree more closely with the p-value obtained under Rao's model. So the p-value obtained with  $G = \hat{\Sigma}_0$  may be suspect, particularly since the estimate of the covariance matrix obtained from the system one oyster lines turned out to be very different from that obtained from the system two oyster lines.

The estimates of the vector of variances obtained respectively from the system one oyster lines and the system two oyster lines were:

[.015, .018, .016, .014, .0096, .026, .035, .051, .040] and

[.025, .024, .041, .020, .012, .0094, .018, .014, .024].

The estimates of the correlation matrix were respectively:

1								
.62	1							
.48	.61	1						
.44	.57	.55	1					
.23	.32	.29	.43	1				
.13	.27	.47	.41	.40	1			
.20	.25	.40	.37	.23	.77	1		
.19	.23	.19	.36	.07	.62	.69	1	
.03	.09	.17	.28	.09	.19	.19	.26	1

and

1								
.72	1							
.74	.73	1						
.52	.59	.61	1					
.54	.61	.58	.68	1				
.47	.54	.49	.55	.63	1			
.24	.22	.13	.30	.27	.31	1		
.35	.35	.35	.50	.57	.61	.47	1	
.25	.22	.15	.29	.32	.43	.45	.59	1

Observations for the system one oyster lines are more variable towards the end of the study than at the beginning, whereas measurements for the system two oyster lines are more variable at the beginning of the study than later on. The two correlation matrices have a roughly similar pattern although the correlations between observations at  $t = 7$  and  $t = 8$  weeks and between observations at  $t = 8$  and  $t = 9$  weeks are particularly low for system two and particularly high for system one.

In view of these differences it may be that using an estimate of  $\Sigma_0$  obtained from one system as the choice of  $G$  in the model for the other system, led to dubious p-values.

### **3.7 Application of the Model to All Seventeen Lines**

The next step was to incorporate the netting and dowing into the model and to test for differences not only between the two systems but

also between the three types of material. Since differences had been found between the system one oyster lines and between the system two oyster lines, each of the seventeen lines was allowed different parameters. After fitting growth curves for each of the seventeen lines, contrasts were examined.

This time only Rao's model was used. Since no estimate of  $\Sigma_0$  was available from a similar but independent experiment in this case, the only reasonable choice of  $G$  in Pothoff and Roy's model would be  $I$ , which could differ substantially from  $\Sigma_0$ .

$Y_0$  (141 x 9) is the data matrix for all 17 lines. The data for system one appears first, then the data for system two.  $P$  is unchanged.

$$A \begin{matrix} (141 \times 17) \end{matrix} = \begin{matrix} 5\{ \\ 8\{ \\ 5\{ \\ 10\{ \end{matrix} \begin{bmatrix} 1 & 0 & . & . & . & . & 0 \\ . & & & & & & \\ . & & & & & & \\ 1 & 0 & . & . & . & . & 0 \\ 0 & 1 & . & . & . & . & 0 \\ . & & & & & & \\ . & & & & & & \\ 0 & 1 & . & . & . & . & 0 \\ 0 & 0 & . & 1 & . & . & 0 \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ 0 & . & . & . & . & 0 & 1 \\ . & & & & & & \\ . & & & & & & \\ 0 & . & . & . & . & 0 & 1 \end{bmatrix}$$

$n = 141$  (number of individuals),  $m = 17$  (number of groups),  
 $q = 9$  (number of time points),  $r = 17$  (rank  $A$ ).

At this stage it was assumed that each set of nine measurements had the same covariance matrix, whether measurements were averages over five as for oyster and dowling, or over ten as for netting.

Suppose we wanted an overall test for differences between the growth curves for the two systems. This could be done by testing the hypothesis:

$$8\underline{\xi}_1 + 8\underline{\xi}_2 + \dots + 8\underline{\xi}_9 - 9\underline{\xi}_{10} - \dots - 9\underline{\xi}_{17} = 0$$

where  $\underline{\xi}_i = [\xi_{i0} \ \xi_{i1} \ \dots \ \xi_{i \ p-1}]$

is the set of  $p$  parameters for group  $i$ . Since the first nine groups correspond to system one lines, and the other eight to system two lines, this is equivalent to testing that the average of the  $\underline{\xi}_i$ 's over the system one lines is the same as the average of the  $\underline{\xi}_i$ 's over the system two lines. A rejection of this hypothesis would indicate a difference between the two systems in the average value of at least one of the parameters. The hypothesis can be written

$$C\underline{\xi}V = 0 \text{ where } C = (8 \ 8 \dots 8 \ -9 \ -9 \ \dots \ -9) \text{ and } V = I \ (p \times p).$$

(1x17)

Other contrasts were examined in order to determine whether differences existed between the three types of material.

First of all, it was determined what degree of polynomial was needed to fit the growth curves for all 17 lines adequately. This was done in the same way as for the oyster lines previously. It was found that quartics were needed. The fitted growth curves obtained for D1, D2, D3 were respectively:

$$y(t) = - .0002 t^4 + .0067 t^3 - .086 t^2 + .578 t + .44$$

$$y(t) = .0005 t^4 - .0143 t^3 + .126 t^2 - .224 t + 1.06$$

$$y(t) = .00001 t^4 + .0009 t^3 - .030 t^2 + .387 t + .51$$

and for N1 and N2:

$$y(t) = .0001 t^4 - .0023 t^3 + .014 t^2 + .100 t + 1.24$$

$$y(t) = .0002 t^4 - .0059 t^3 + .058 t^2 - .113 t + 1.62$$

Pictures of the curves for the dowling lines appear in Figure 3 and for the netting lines in Figure 4. The curves for the netting lines are almost linear. A test of the hypothesis of linearity for these curves alone yielded a p-value as large as 7%. We have already ascertained that quadratics are adequate for the oyster lines. Clearly it is because of the dowling lines, and because of D2 in particular, that quartics are needed. The coefficient of  $t^4$  is much larger for D2 than for any other line and the growth curve for D2 looks remarkably different from any of the other growth curves (see Figure 2). Before discussing this, results of the tests comparing the curves for the two systems and for the three material types will be presented:

1. The growth curves for the two systems were compared by testing

$$C\xi V = 0 \text{ where } C = \begin{pmatrix} 8 & 8 & \dots & 8 & -9 & -9 & \dots & -9 \end{pmatrix} \text{ and } V = I \text{ (5x5).}$$

(1x17)

$$\text{Largest root of } S_h S_e^{-1} = .732,$$

$$\text{Refer 16.9 to } F_{5,116}, p < .001.$$

Figure 3. Growth Curves for the Dowling Lines

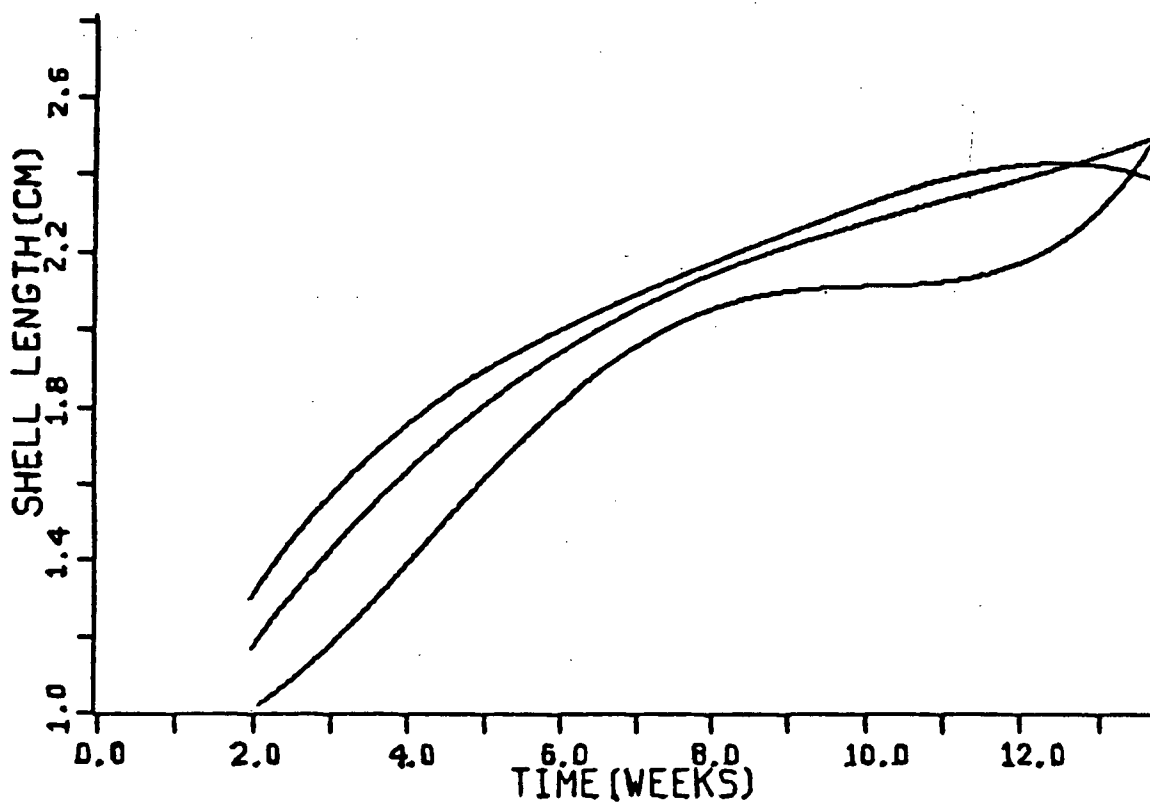
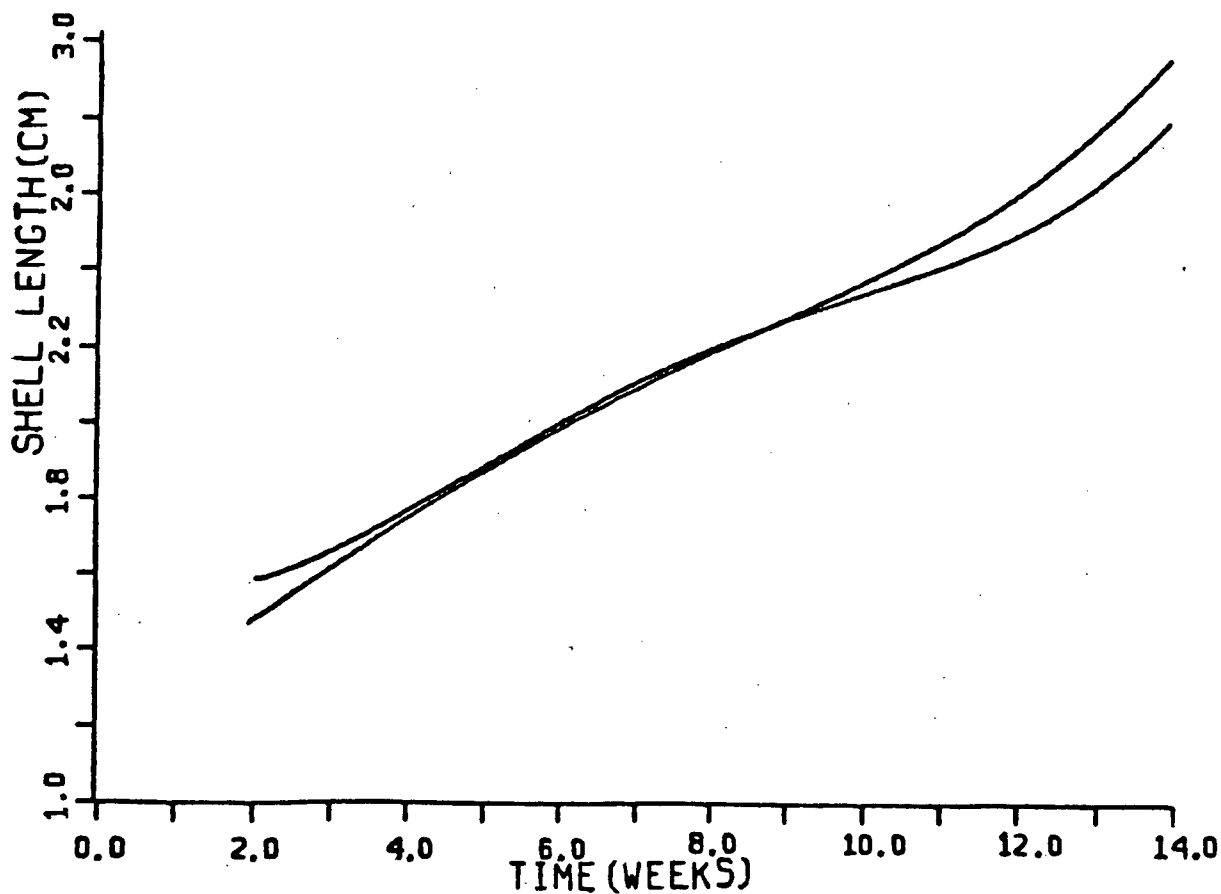


Figure 4. Growth Curves for the Netting Lines



2. The growth curves for the oyster and dowling lines were compared by testing  $C\xi V = 0$  where  $C = (1 \ 1 \ 1 \ 1 \ 1 \ 1 \ -4 \ -4 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ -4 \ 0)$  and  $V = I \ (5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = 1.121$ ,

Refer 25.9 to  $F_{5,116}$ ,  $p < .001$ .

3. The growth curves for the oyster and netting lines were compared by testing  $C\xi V = 0$  where  $C = (1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ -6 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ -6)$  and  $V = I \ (5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = .426$ ,

Refer 9.9 to  $F_{5,116}$ ,  $p < .001$ .

4. The growth curves for the dowling and netting lines were compared by testing  $C\xi V = 0$  where  $C = (0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 2 \ 2 \ -3 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 2 \ -3)$  and  $V = I \ (5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = .524$ .

Refer 12 to  $F_{5,116}$ ,  $p < .001$ .

The same analyses were then repeated this time allowing the curves to differ by an additive constant. So this time  $\xi_0$  was not included; for each test  $C$  remained unchanged, but  $V$  was replaced by

$$\begin{vmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}.$$

## Results

### 1. The two systems

Largest root of  $S_h S_e^{-1} = .412$

Refer 12.2 to  $F_{4,119}$

$p < .001$  .

### 2. Oyster and dowling

Largest root of  $S_h S_e^{-1} = .730$

Refer 21.7 to  $F_{4,119}$

$p < .001$  .

### 3. Oyster and netting

Largest root of  $S_h S_e^{-1} = .357$

Refer 10.7 to  $F_{4,119}$

$p < .001$  .

### 4. Dowling and netting

Largest root of  $S_h S_e^{-1} = .471$

Refer 13.9 to  $F_{4,119}$

$p < .001$  .

## Discussion

Differences are found both between the two systems and between the three types of material. This is in contrast to what was found for the



survival distributions, namely that only system was an important factor.

With such a large amount of data it is not surprising that differences are found everywhere. The relative size of the test statistics indicates that the biggest differences lie between dowling and the other two types of material. This was to be expected in view of the unusual growth curve for D2.

For the comparisons of the two systems and of oyster and dowling lines the p-values were much larger when the curves were allowed to differ by an additive constant. On the other hand for the other two comparisons the p-values did not change substantially when the curves were allowed to differ by a constant. So the difference between netting and dowling and between netting and oyster is (almost) solely attributable to the different shapes of the growth curves. The difference between the two systems and between oyster and dowling is due partly to the different shapes of the curves and partly to different locations. We conclude that barnacles on system two are generally bigger than those on system one and that barnacles on oyster lines are generally bigger than those on dowling lines.

The smoothness of the growth curves for the netting lines may reflect the averaging over ten instead of five, rather than a growth characteristic.

The average size of barnacles on netting lines appears to increase at an almost constant rate except towards the end of the study when it increases a little faster, whereas the average size of barnacles on dowling lines is particularly small initially, increases rapidly to

start with and less rapidly later on. The behavior of the average size of barnacles on oyster lines lies somewhere between these two extremes.

The average size of barnacles on D2 is exceptionally small at  $t = 2$  weeks, increases fairly rapidly until  $t = 8$  weeks, appears to remain the same until  $t = 12$  weeks and then to increase extremely rapidly. This could be a growth characteristic, but since this behavior is so different, it may also be that particularly small barnacles were sampled at  $t = 8$  and 9 weeks and particularly large barnacles were sampled at  $t = 14$  weeks.

### **3.8 Extension of the Model to Allow Different Covariance Matrices for Different Items**

An estimate of  $\Sigma_0$  was obtained separately from the oyster lines, the dowling lines and the netting lines. Not surprisingly the estimate obtained from the netting lines was considerably smaller than the other estimates since measurements from items on netting lines are averages over ten instead of five. It would seem more reasonable to assume that the covariance matrix for items on oyster or dowling lines is  $(1/5)\Sigma$  for some  $\Sigma$ , and  $(1/10)\Sigma$  for items on netting lines. The growth curve models can be extended to the case where the covariance matrices for different individuals are assumed to be known multiples of each other (Ito, 1968). Suppose that there are  $N$  individuals who are assumed to have covariance matrices

$$\theta_1 \Sigma, \theta_2 \Sigma, \theta_3 \Sigma, \dots, \theta_N \Sigma.$$

where the  $\theta_i$  are scalars and  $\Sigma$  is fixed but unknown. Suppose the

individuals are divided into groups with  $n_j$  in the  $j$ th group.

Individuals in the same group are assumed to have the same covariance matrix, so  $\theta_1 = \theta_2 = \dots = \theta_{n_1}$  and similarly for the other groups.

Define a diagonal matrix of weights

$$W = \begin{vmatrix} 1/\theta_1 & & & \\ & 1/\theta_2 & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & 1/\theta_N \end{vmatrix}$$

(NxN)

Instead of choosing  $\xi$  to minimise  $(X - A\xi)' (X - A\xi)$  it is now chosen to minimise  $(X - A\xi)' W (X - A\xi)$  (weighted least squares). So more weight is given to individuals with smaller covariance matrix. It follows

that  $\hat{\xi} = (A_1' W A_1)^{-1} A_1' W X$ .

Tests of  $C\xi V = 0$  are based on

$$S_h = (C_1 \hat{\xi} V)' (C_1 R C_1)^{-1} (C_1 \hat{\xi} V)$$

where  $R = (A_1' W A_1)^{-1} + (A_1' W A_1)^{-1} A_1' W X B X' W A_1 (A_1' W A_1)^{-1}$

and  $S_e = V'(P S^{-1} P')^{-1} V$ .

If  $W = I$  (this means equal weight is given to all observations) we obtain the same expression for  $R$  and  $\hat{\xi}$  and hence for  $S_h$  as previously.

The estimate of  $\Sigma$  is given by  $S$ , where

$$S = \frac{1}{\theta_1} S_1 + \frac{1}{\theta_{n_1+1}} S_2 + \frac{1}{\theta_{n_1+n_2+1}} S_3 + \dots$$

and  $S_t$  is the estimate of the covariance matrix for group  $t$ , namely

$$S_t = Y_0' [I - A_1 (A_1' A_1)^{-1} A_1'] Y_0 \text{ where } Y_0 \text{ is the data matrix for group } t$$

and  $A$ , the design matrix for group  $t$ , is a  $(n_t \times 1)$  vector of 1's.

Individuals in group 2 have covariance matrix  $\theta_{n_1+1} \Sigma$ , those in group 3

have covariance matrix  $\theta_{n_1+n_2+1} \Sigma$  and so on. If  $\theta_1 = \theta_2 = \dots = \theta_N = 1$

(all individuals have the same covariance matrix) then  $S$  will be as

for the standard model, namely  $Y'_0 [I - A_1 (A'_1 A_1)^{-1} A'_1] Y_0$  where  $Y_0$  and  $A$

are the data and design matrices for the complete set of data. The

tests based on  $S_h$  and  $S_e$  are then as before. Quartics were again

fitted and all the tests of the previous section were repeated with the

adjusted  $S_h$  and  $S_e$ .

In our case  $S = \frac{1}{2} [\text{covariance matrix for oysters}] +$

$\frac{1}{2} [\text{covariance matrix for dowling}] + [\text{covariance matrix for netting}]$

and  $\theta_i = 2$ ,  $i = 1, \dots, 60$  and  $69, \dots, 131$  (items on oyster and dowling

lines),

$\theta_i = 1$ ,  $i = 61, \dots, 68$  and  $132, \dots, 141$  (items on netting lines).

$A$ ,  $Y_0$  and  $P$  are unchanged.

## Results

The same tests as in the previous section, to compare the growth curves for various pairs of subgroups were carried out:

### 1. The two systems

Largest root of  $S_h S_e^{-1} = .595$

Refer 13.9 to  $F_{5,116}$ ,  $p < .001$

2. Oyster and Dowling

Largest root of  $S_h S_e^{-1} = .902$

Refer 20.9 to  $F_{5,116}$ ,  $p < .001$

3. Oyster and Netting

Largest root of  $S_h S_e^{-1} = .610$

Refer 14.2 to  $F_{5,116}$ ,  $p < .001$

4. Netting and Dowling

Largest root of  $S_h S_e^{-1} = .640$

Refer 14.8 to  $F_{5,116}$ ,  $p < .001$

Comments

The estimate of  $\xi$  is unaffected by including W in the analysis, since within each group (line) the same weight is given to all items (the same covariance matrix is assumed for individuals in the same group).

However the p-values of the tests are affected. In comparisons involving netting, the p-value obtained in the original unweighted analysis was larger. This is because the estimate of the covariance was too big because it was obtained from all items, whether from oyster, netting or dowling lines, and these were all assumed to have the same  $\Sigma_0$ . For comparisons not involving netting the original p value was

smaller. This is because the estimate of the covariance matrix was too small because it was obtained not only from items on oyster and dowling lines but also from items on netting lines. The p-values obtained in the weighted analysis should be expected to be more reliable.

The main conclusions remain the same however, namely that differences exist both between the two systems and between the three types of materials with the biggest difference being between the oyster and dowling lines.

### 3.9 Collapse Over Lines Within Each System-Type Combination

Although differences had been found between lines of the same type within a system, (particularly within system two), another analysis was then carried out in which these differences were ignored. Ignoring lines we get a two-factor situation as previously described. The number of items in each category is given below:

Type	System 1	System 2
Oyster	38	50
Netting	22	13
Dowling	8	10

The analysis was unweighted - the same covariance matrix was assumed for all individuals (items). It was assumed that the growth curve for an item on the jth type of line in the ith system was

$$(\xi_{0i} + \xi_{1i} t + \xi_{2i} t^2 + \dots + \xi_{p-1 i} t^{p-1})$$

$$+ (\mu_{0j} + \mu_{1j} t + \dots + \mu_{p-1j} t^{p-1})$$

So  $E[Y_0] = A \xi P$ ,

where  $A (141 \times 5) = 38$  {

1	0	0	1	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0	1	0	1	0
.	.	.	.	.
.	.	.	.	.
0	0	1	1	0
.	.	.	.	.
.	.	.	.	.
1	0	0	0	1
.	.	.	.	.
.	.	.	.	.
0	1	0	0	1
.	.	.	.	.
.	.	.	.	.
0	0	1	0	1
.	.	.	.	.
.	.	.	.	.

22 {

8 {

50 {

13 {

10 {

$$\xi(5xp) = \begin{vmatrix} \mu_{01} & \mu_{11} & \dots & \mu_{p-1 1} \\ \mu_{02} & \mu_{12} & \dots & \mu_{p-1 2} \\ \mu_{03} & \mu_{13} & \dots & \mu_{p-1 3} \\ \xi_{01} & \xi_{11} & \dots & \xi_{p-1 1} \\ \xi_{02} & \xi_{12} & \dots & \xi_{p-1 2} \end{vmatrix}$$

and  $Y_0$  and  $P$  are as before.

Quartics were fitted and the usual hypotheses were tested, namely that the growth curves for various pairs of subgroups were identical:

1. The two systems

Test  $C \xi V = 0$  where  $C = (0 \ 0 \ 0 \ 1 \ -1)$  and  $V = I(5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = .752$ .

Refer 19.2 to  $F_{5,128}$ ,  $p < .001$

2. Oyster and Dowling

Test  $C \xi V = 0$  where  $C = (1 \ -1 \ 0 \ 0 \ 0)$  and  $V = I(5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = 4.561$

Refer 116.7 to  $F_{5,128}$ ,  $p \ll .001$

3. Oyster and Netting

Test  $C \xi V = 0$  where  $C = (1 \ 0 \ -1 \ 0 \ 0)$  and  $V = I(5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = .874$

Refer 22.4 to  $F_{5,128}$ ,  $p < .001$

4. Netting and Dowling

Test  $C \xi V = 0$  where  $C = (0 \ 1 \ -1 \ 0 \ 0)$  and  $V = I(5 \times 5)$ .

Largest root of  $S_h S_e^{-1} = 5.022$

Refer 128.5 to  $F_{5,128}$ ,  $p \ll .001$

Again differences are found everywhere with the biggest differences being between dowling and each of the other two types. The smallest difference is between the two systems. However the p-values for the two comparisons involving dowling are suspiciously small and this method of collapsing over lines is suspect.



### 3.10 Missing Data

Another generalisation of the model could also have been used, namely the generalisation suggested by Kleinbaum (1973) to allow missing data. In this way the assumption that observations are taken at the same times for all individuals (items) could have been relaxed. Under this model it is assumed that data is missing at random - whether or not an observation is missing is unrelated to the value it would have taken. Kleinbaum's generalised model proceeds as follows: The  $n$  individuals are divided into  $u$  disjoint groups  $S_1, S_2, \dots, S_u$  such that individuals in the same group  $S_j$  have measurements taken at the same  $q_j$  time points. The number of individuals in  $S_j$  is  $n_j$ . Individuals in different groups may not have measurements taken at the same time points but may have measurements taken at the same number of time points ( $q_j = q_{j'}$ ).

Let  $q$  be the total number of time points. Then for each group  $S_j$  we have a design matrix  $A_j$  ( $n_j \times m$ ) (since within each group  $S_j$ , individuals are subdivided according to which of the  $m$  "treatment" groups they belong to). Also for each group  $S_j$  we have an incidence matrix  $B_j$  ( $q \times q_j$ ) of 0's and 1's indicating the positions of missing observations for individuals in  $S_j$ . If  $Y_j$  ( $n_j \times q_j$ ) is the data matrix for the individuals in  $S_j$  and  $\Sigma_0$  is the covariance matrix for a complete set of  $q$  observations on any individual (assumed the same for every individual), then for each  $j = 1, 2, \dots, u$  we have:

$$E[Y_j] = A_j \xi P B_j$$

and for each row of  $Y_j$  the covariance matrix is  $B_j' \Sigma_0 B_j$ , where  $\xi$  and  $P$  are as for the growth curve models for complete data described in Section 3.3.

As before the rows of  $Y_j$  are assumed to be mutually independent and each row is assumed to have a multivariate normal distribution. Under this model it is not possible to find the maximum likelihood estimates of  $\xi$  and  $\Sigma$  in closed form, but Kleinbaum finds some best asymptotically normal (BAN) estimators which have the same large sample properties as the maximum likelihood estimates.

Let  $\xi^*$  be the column vector formed by putting the columns of  $\xi$  underneath each other and let  $y_j$  be the vector formed by putting the columns of  $Y_j$  underneath each other. Then a BAN estimator of  $\xi^*$  (which Kleinbaum finds by writing the model in the form of a general univariate weighted least squares model) turns out to be:

$$\hat{\xi}^* = \left[ \sum_{j=1}^u P B_j (B_j' \hat{\Sigma}_0 B_j)^{-1} B_j P' \otimes A_j' A_j \right]^{-1} \sum_{j=1}^u [P B_j (B_j' \hat{\Sigma}_0 B_j)^{-1} \otimes A_j'] y_j$$

where  $M^-$  indicates the generalized inverse of the matrix  $M$  and  $\hat{\Sigma}_0$ , an unbiased and consistent estimator of  $\Sigma_0$ , is found as follows: The (r,s) element  $\sigma_{rs}$  of  $\hat{\Sigma}_0$  is the usual pooled estimate using only

those individuals for which measurements were obtained at both  $t_r$  and  $t_s$ . In order to test a hypothesis of the form  $H'\xi^* = 0$ , where  $H$  (mp x w) is of full rank w and  $H'\xi^*$  is estimable (which means that it has an unbiased estimator linear in y where y is the vector formed by putting

the columns of Y underneath each other), the test statistic is given by

$$W_n = (H' \hat{\xi}^*)' [H' \left[ \sum_{j=1}^u P B_j (B_j' \hat{\Sigma}_0 B_j)^{-1} B_j P' \otimes A_j' A_j \right]^{-1} H]^{-1} (H' \hat{\xi}^*)$$

where  $\hat{\xi}^*$  is any BAN estimator of  $\xi^*$  and  $\hat{\Sigma}_0$  is any consistent positive definite estimator of  $\Sigma_0$ . In particular the estimators obtained by Kleinbaum may be used. Under  $H_0$ ,  $W_n$  is asymptotically distributed as  $X_w^2$ .  $W_n$  is the analogue of  $S_h S_e^{-1}$  in Pothoff and Roy's model described in Section 3.3. An alternative statistic  $W_n^*$  may be used which is the analogue of  $S_h S_e^{-1}$  in Rao's model.  $W_n^*$  is obtained by replacing  $(A_j' A_j)$  in  $W_n$  by a more complicated matrix  $Q_j$  whose generalized inverse is analogous to  $R$  of Rao's model.

#### Application of the Model to our Data

This generalisation might have been appropriate for our data set, because some data was neglected in order to satisfy the assumption that observations on all individuals were taken at the same times. Namely the data taken at  $t = 1, 10$  and  $11$  weeks was ignored.

Under Kleinbaum's model this data could have been included: The items would have been divided into groups first according to the times at which data was taken on them. For example items for which data was available at 2, 3, ..., 9, 10, 14, 17 would form one group, those for which data was available at 2, 3, ..., 9, 11, 14, 17 would form another group and so on. Within each group items would be divided according to line. Thus the  $A_j$ 's would be determined. Then polynomials would have

been fitted as before and the parameters for different lines compared.

However since the amount of data neglected was so small compared to the total amount of data, and since application of the model would have been quite laborious it was not considered worthwhile to pursue this model.

### 3.11 Growth Curves for Individual Barnacles

In all the previous analyses relating to the growth data, after averaging over the sample of five (ten), there was only one observation at each time for each item. For this reason it was not possible to fit a different growth curve for each item. So our covariance matrix contained two components of variability - barnacle-to-barnacle variability within items and item to item variability. It would be nice to fit the average growth curve for individuals on each item and then to compare the growth curves for the different items within a line. Whether this is possible with the limited data that is available will be investigated in this section.

Consider just the first oyster line. Ideally a data matrix  $Y_1$  would be available, each row of which would correspond to one barnacle. There would be five rows corresponding to each item on the line. In fact all that is available is a data matrix  $Y_2$ . In  $Y_2$ , five rows correspond to each item, but within each set of five rows, successive observations in a row may or may not correspond to the same barnacle. Furthermore observations at different time points in different rows may

correspond to the same barnacle, so the rows are dependent. However, the five observations at any one time are independent.

But  $E[Y_1] = E[Y_2] = A\xi P$ . So we can use our data matrix  $Y_2$  just as we would have used  $Y_1$  to obtain a least squares estimate of  $\xi$ . This estimate involves only the average of the observations at each time and it doesn't matter whether or not the observations at successive times correspond to the same barnacles - our estimate is still valid. Of course if we were to assume that all items have the same growth curve, we would obtain the same growth curve for the oyster line as previously when we first averaged within each item. But by choosing a different design matrix  $A$ , we can now obtain a different growth curve for each item.

Testing for differences between the growth curves is not as straightforward as fitting the growth curves: in order to test hypotheses relating to the parameters we need to find matrices  $S_h$  and  $S_e$ . These involve  $\hat{\xi}$  and  $S$ , an estimate of the covariance matrix  $\Sigma_0$  for the set of observations on an individual barnacle. We have already obtained  $\hat{\xi}$ , but  $S$  is more difficult. Since each row of  $Y_2$  does not necessarily correspond to the same barnacle,  $Y_2'[I - A_1(A_1' A_1)^{-1} A_1'] Y_2$  does not provide an estimate of  $\Sigma_0$ . However if a strong assumption is made about the form of  $\Sigma_0$ , it is possible to obtain an estimate of it.

### Obtaining an Estimate of $\Sigma_0$

If we assume that variance is constant over time and that the correlation between observations  $d$  weeks apart on an individual is  $\rho^d$  then  $\Sigma_0$  is proportional to:

1	$\rho$	$\rho^2$	$\rho^3$	$\rho^4$	$\rho^5$	$\rho^6$	$\rho^7$	$\rho^{12}$
	1	$\rho$	$\rho^2$					
		1	$\rho$	$\rho^2$				
			1	$\rho$	$\rho^2$			
				1	$\rho$	$\rho^2$		
					1	$\rho$	$\rho^2$	$\rho^7$
						1	$\rho$	$\rho^6$
							1	$\rho^5$
								1

Under this strong assumption all that is needed is an estimate of  $\rho$ .

This is obtained as follows:

A number of items were excluded from the previous analyses because towards the end of the study, fewer than five barnacles were still alive on them. It was assumed that these barnacles were representative of all the barnacles and an estimate of  $\rho$  was obtained using only these barnacles. Suppose that on a particular item, four barnacles remained for the last few time points. Then clearly we would be sampling the same four barnacles at each of these times. Furthermore it was usually clear which observations at successive times corresponded to the same barnacles (since barnacles cannot get smaller). If this wasn't obvious

it was assumed that the smallest observation at the first time corresponded to the smallest observation at the next time, and so on.

Items with only one barnacle left could not be used, as replication was needed. Eleven items (all from the system one oyster lines) were used. Longitudinal data was available at  $t = 11, 14$  and  $17$  weeks on  $2, 3$  or  $4$  barnacles on each of these items; the total number of barnacles involved was  $34$ . Using this data, the maximum likelihood estimate of  $\rho$  was obtained. Koopmans (1942) describes how to find the maximum likelihood estimate of  $\rho$  under our model of serial correlation. Since our longitudinal data was taken at three week intervals we let  $\rho' = \rho^3$ . Our assumption about  $\Sigma_0$  implies that

$$(y_t - \mu_t) = \rho'(y_{t-1} - \mu_{t-1}) + Z_t$$

where the  $Z_t$  are independently distributed as  $N(0, \sigma^2)$ ,  $y_t$  and  $y_{t-1}$  are observations on the same individual respectively at  $t$  and  $t-1$ , and  $\mu_t = E[y_t]$ . It follows that the variance of the  $y$ 's is  $\sigma^2/(1 - \rho'^2)$ .

The likelihood of obtaining the  $y$ 's is,

$$L = \frac{(1 - \rho'^2)^{N/2}}{(2\pi\sigma^2)^{NT/2}} \exp \left[ - (A - 2\rho'B + (1 + \rho'^2)C)/2\sigma^2 \right]$$

where  $T$  is the total number of time points,  $N$  is the total number of individuals and

$$A = \sum_i \sum_j [(y_1^{ij} - \mu_1^i)^2 + (y_T^{ij} - \mu_T^i)^2],$$

$$B = \sum_i \sum_j [(y_1^{ij} - \mu_1^i)(y_2^{ij} - \mu_2^i) + \dots + (y_{T-1}^{ij} - \mu_{T-1}^i)(y_T^{ij} - \mu_T^i)],$$

$$C = \sum_i \sum_j [(y_2^{ij} - \mu_2^i)^2 + \dots + (y_{T-1}^{ij} - \mu_{T-1}^i)^2],$$

where  $y_t^{ij}$  is the observation at time  $t$  on the  $j$ th barnacle on the  $i$ th item and  $\mu_t^i = E[y_t^{ij}]$ . Maximising this likelihood over the  $\mu_t^i$  and  $\rho'$  leads to an estimate of  $\rho'$  as the root of a cubic equation (Koopmans, 1942).

### Results

The estimate of  $\rho'$  turned out to be .899. So  $\hat{\rho} = .965$ . The correlation between observations on the same individual at successive times is seen to be very high. It is possible to obtain estimates of the barnacle-to-barnacle variance within items at each time and thus check our assumption of constant variance over time. The variance at time  $t$  is estimated by

$$\sum_i \sum_j \frac{(y_t^{ij} - \bar{y}_t^i)^2}{N - 1}$$

where  $l$  is the number of items.

The estimates turned out to be .054, .069, .056, .054, .048, .049, .060, .043 and .053 suggesting that our assumption of constant variance over time was a fairly weak one. Combining these estimates would then lead to an estimate of  $\sigma$  and thus to an estimate of  $\Sigma_0$  of the assumed serial correlation form. Although this is probably the best estimate of  $\Sigma_0$  that can realistically be obtained from our limited data it is unlikely to be a good estimate for several reasons: firstly a strong assumption was made about the form of  $\Sigma_0$  and secondly the estimate was obtained using only a small number of barnacles and using observations only at the last three time points.



### Testing Hypotheses of the Form $C\xi V = 0$

Having obtained an estimate  $\hat{\Sigma}_0$  of  $\Sigma_0$ , is it possible to test hypotheses of the form  $C\xi V = 0$  by substituting  $\hat{\Sigma}_0$  for  $S$  and  $\hat{\xi}$  obtained from  $Y_2$  for  $\xi$  in the expressions for  $S_h$  and  $S_e$ ? This is not clear.

The tests of hypotheses which have been discussed rely on the fact that if  $\hat{\mu} \sim N(\mu, c\Sigma)$  where  $c$  is constant, and  $D \sim W_p(k, \Sigma)$ , the Wishart distribution with  $k$  degrees of freedom, then the distribution of  $\hat{\mu}' D^{-1} \hat{\mu}$  is known (namely Hotelling's  $T^2 = \frac{k}{c} \hat{\mu}' D^{-1} \hat{\mu}$ ). Suppose an estimate of  $\xi$  is obtained from the data matrix  $Y_1$  (the true longitudinal data), namely

$$\hat{\xi} = (A_1' A_1)^{-1} A_1' Y_1 P'(P P')^{-1}$$

where we have taken  $G = I$  in Pothoff and Roy's model. Then the covariance matrix of  $\hat{\xi}$  is

$$(P P')^{-1} P(A_1' A_1)^{-1} \Sigma_0 P'(P P')^{-1} = (A_1' A_1)^{-1} \Sigma$$

where  $\Sigma$  is the covariance matrix for one row of  $X = Y P' (P P')^{-1}$ .

Further

$$\hat{\xi} \sim N(\xi, (A_1' A_1)^{-1} \Sigma) \quad (1)$$

and

$$X'[I - A_1(A_1' A_1)^{-1} A_1'] X \sim W(k, \Sigma). \quad (2)$$

Together (1) and (2) imply that the distribution of  $S_h S_e^{-1}$  is known and this forms the basis for the tests of  $C\xi V = 0$ .

Now suppose that the estimate of  $\xi$  is obtained instead from our data matrix  $Y_2$  (let this estimate be denoted  $\hat{\xi}_1$ ) and that

$S = Y'_0 [I - A_1 (A'_1 A_1)^{-1} A'_1] Y_0$  is replaced by our estimate  $\hat{\Sigma}_0$  of  $\Sigma_0$ .

Then there are two reasons why the same distribution theory no longer applies:

1. Although  $\hat{\xi}_1$  still has a multivariate normal distribution with mean  $\xi$ , its covariance matrix is not the same as that of  $\hat{\xi}$  and in particular is not proportional to  $\Sigma$ .

2.  $\hat{\Sigma}_0$  does not have a Wishart distribution as  $S$  does because of the special form that was assumed for it.

So although it is possible to obtain growth curves for individual barnacles without longitudinal data, it does not seem possible to test for differences.

### 3.12 Discussion

The growth curve models that were applied the previous sections to our longitudinal data for items required a number of assumptions. As mentioned in Section 3.4 the only assumption that was likely to cause a problem was the assumption of equal covariance matrices for every individual (item). There were two reasons why this might be a strong assumption even if the covariance matrices were assumed to be the same for every individual barnacle:

1. Our observations were averages over five barnacles in some cases and over ten barnacles in other cases.

2. At each time the population from which the barnacles were sampled was of a different size for each different item.

In Section 3.8 the analysis was modified to allow for the fact that observations were averages over sometimes five and sometimes ten barnacles. New p-values were obtained which were slightly different from those obtained previously. However the broad conclusions remained the same.

It is not so straightforward to accommodate the different population sizes, especially since the number of barnacles alive at each time on each item is not known (we know only the total number on each line). In Section 3.6 estimates of the covariance matrices obtained respectively from the system one oyster lines and from the system two oyster lines were presented. The two estimates did indeed show somewhat different patterns which may be partly due to the different population sizes. In view of this difficulty, the p-values obtained in Sections 3.5, 3.6 and 3.7 should not be considered completely reliable. However the p-values that we obtained were so small, that even if the error involved were quite large, we would still reach the same conclusions.

In Sections 3.5, 3.6 and 3.7, growth curves for items were fitted. In these sections each data point was an average over a number of barnacles alive at that time. So our growth curves represented the natural population. In Section 3.11 we discussed fitting growth curves for individual barnacles. But again our estimate of  $\xi$  was obtained from data on the barnacles which happened to be alive at any particular time. So again the curves represented the natural population.

Ideally we would like to fit growth curves that would represent the growth of a typical individual barnacle between  $t = 2$  and 17 weeks. It

is not clear that this is possible with the limited data available.

Ideally the following information would be available for each barnacle:

- (a) Growth data at each time point until  $t = 17$  weeks or until its death whichever is the earlier.
- (b) Time of death if this is before  $t = 17$  weeks or knowledge that it survived past  $t = 17$ .
- (c) Cause of death - food shortage or predator.
- (d) Number of the item and line to which it is attached.

If this were the case several approaches would be possible depending on the objectives:

1. Suppose we wanted to fit growth curves for individual barnacles while accounting for the censoring. The deaths would affect the analysis only in one way, namely the growth data for each individual would be truncated at its death time if this is before  $t = 17$  weeks; we would have growth data with missing observations.

Kleinbaum's model is sufficiently general to allow this pattern of missing data, so this model could be used to fit growth curves and to test for differences between various groups. The growth curves that we would obtain would represent not the natural population, but the growth experience of a typical barnacle between  $t = 2$  and 17 weeks.

2. If we were interested instead in whether survival was related to growth we could think of our data as survival data with a time dependent covariate. We would have type I censoring in the survival data; for barnacles which survive past  $t = 17$  weeks, the exact lifetime

would not be known. The time dependent covariate could be shell length.

Time dependent covariates can be incorporated into Cox's proportional hazards model, which is described in Section 2.1. The hazard function would then be  $\lambda(t; z(t)) = \lambda_0(t) e^{z(t)\beta}$  where in our case:

$\lambda_0(t)$  is an unknown baseline hazard function,

$\beta$  is an unknown parameter,

$z(t)$  is our time dependent covariate - shell length at time  $t$ .

In this context a test of  $\beta = 0$  would be a test that growth characteristics and survival time are not related. The method of testing  $H_0: \beta = 0$  is described in Section 2.1. A rejection of this hypothesis would suggest some dependency between growth characteristics and survival time, for example a tendency for larger barnacles to live longer.

3. There are two possible causes of death for the barnacles, namely they may be eaten by predators or they may die due to food shortage. Instead of testing whether growth is related to survival time, it may be of interest to test separately whether each cause of death is related to growth. Then we would be able to address questions such as "do the bigger barnacles tend to be eaten first by the predators?" and "is it the smaller barnacles which tend to die of food shortage?"

This can be done in the context of Cox's proportional hazards model by finding the hazard function specific to each cause of death. Growth

is again included as a time dependent covariate but this time  $\beta$  may vary over the two types of failure. So estimates of  $\beta_1$  and  $\beta_2$  would be obtained and a test of  $\beta_1 = 0$  would be a test that the  $i$ th cause of death is unrelated to growth. To obtain an estimate of  $\beta_1$ , all failures other than those of type  $i$  are treated as censored observations, and the usual maximum likelihood methods are used.

Unfortunately we do not have all the information that we would like: at each time point we do not have growth data on all barnacles, only on five from each item. Furthermore individual barnacles cannot be identified so it is not known whether or not successive observations correspond to the same barnacles, nor which deaths correspond to which growth measurements. The cause of death is not known. Survival data is available only for the lines not for the items within the lines. For these reasons we cannot pursue any of the approaches described above. However, the growth experience of a typical barnacle between  $t = 2$  and 17 weeks on a particular line will be well represented by the growth curve that was fitted for that line in Section 3.7, provided that the probability of a barnacle dying is unrelated to its size. This can be explained as follows: if the barnacles that die between two time points are neither particularly big or particularly small, then the change in average size of living barnacles over this time interval will be due solely to barnacles growing, not to the changing population.

With our limited data it is not easy to ascertain whether the probability of dying is related to size: since survival data is available only for the lines, growth data was averaged over items within

each line. So for each of the 17 lines average shell length and the percentage of deaths were available at each time.

Using all 17 lines a significant correlation was found between size at week 5 and the percentage dying between weeks 5 and 6. On those lines where average size at week 5 was large, a smaller percentage tended to die between weeks 5 and 6. However we cannot conclude from this that within a particular line the smaller barnacles tend to die first. The correlation may be due to a third underlying factor related to both survival and growth. It may reflect, for example, a system or treatment (type of material) effect - barnacles on system two tend to start dying later and furthermore they are generally bigger than barnacles on system one. To eliminate a possible system or treatment effect, we should use, for example, just the system one oyster lines. However there are only six of these so a correlation obtained just from these lines wouldn't be very meaningful. If survival data were available for every item, we would have average size and percentage dying at each time for 60 items from system one oyster lines. In this case it would even be possible to eliminate a possible line effect by using just the items from one line. Then a significant correlation at a particular time between average size and percentage dying would certainly suggest dependency between the probability of dying and size.

From a biological point of view the experimenter expected that size and lifetime would be unrelated. Without detailed survival data we can only assume that this is true.

One other point is of interest: in Section 2.5 it was noted that there was a strong correlation between the initial number on a line and

the estimates of  $\beta$  and  $\alpha$  for that line. The correlation suggested that on lines with fewer barnacles initially (e.g. system one oyster lines), barnacles tended to start dying sooner and at a faster rate.

We now have the additional observation that barnacles on system one are generally smaller initially than those on system two. We could speculate from these observations that system two is a more ideal site for the barnacles which is why more barnacles become attached to system two. But maybe there isn't room for all the barnacles on system 2 and they have to compete for a place. The bigger healthier barnacles may have a better chance of finding a place on system 2, which would explain why the barnacles on system two are initially bigger. The barnacles on system one tend to start dying earlier and at a faster rate. Also barnacles on the dowling lines are particularly small initially and start dying early and at a fast rate. The fact that the population of barnacles on system one and on dowling lines diminishes quickly may be due to the fact that the barnacles here are small to start with or may be attributable to a poorer environment - it is difficult to say which.

Similarly it is difficult to know whether to attribute differences in the growth curves to the different environments or to the initial differences in the barnacles.



## CONCLUSION

A data set provided by Mr. H. Goldberg relating to the survival and growth of barnacles was examined. Nonparametric tests, namely the logrank and the Wilcoxon, indicated that differences existed between the survival distributions even of lines of the same material type within the same system. The tests suggested that system had more influence than material type on the survival distribution. The exponential model was fitted for the survival distributions of each of the 17 lines and the estimated slopes,  $\hat{\beta}$  were compared.

A random effects model was developed for the  $\beta$  values, initially incorporating only the oyster lines but subsequently including all 17 lines. Maximum likelihood estimates were obtained of the true underlying  $\beta$  value for lines of each type within each system. A comparison of these estimates suggested that slope was dependent on system but not on material type. Again system was found to be the more important factor.

The growth data was less straightforward to examine as we did not have longitudinal data. We averaged over the set of measurements obtained at each time from each item and treated the resulting data set as longitudinal data for the items. Using this set of averages, polynomial growth curves for items were fitted to each line using both Pothoff and Roy's model and Rao's model. Quadratics were adequate except in the case of the second dowling line. A comparison of the parameters indicated that differences existed both between systems and

between material types, with the biggest differences being between dowling and the other two types of material. The p-values may be somewhat unreliable as the strong assumption of equal covariance matrices for every individual (item) had to be made. Although, we didn't have longitudinal data for barnacles it was found to be possible to fit the average growth curve for individual barnacles on each item but not to test for differences between the curves obtained.

In conclusion, although we had a large amount of growth data, a smaller amount of longitudinal data would have been more useful.

## BIBLIOGRAPHY

- Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34, 187-202.
- Grizzle, J. and Allen, D.M. (1969). Analysis of growth and dose response curves. *Biometrics* 25, 357-381.
- Ito, K. (1968). On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. In: Krishnaiah, P.R. ed.; *Multivariate Analysis II*. New York: Academic Press, 87-120.
- Kalbfleisch, J.G. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York: Wiley.
- Khatri, C.G. (1966). A note on a manova model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics* 18, 75-86.
- Kleinbaum, D.G. (1973). A generalization of the growth curve model which allows missing data. *Journal of Multivariate Analysis* 3, 117-124.
- Koopmans, T. (1942). Serial correlation and quadratic forms in normal variables. *Annals of Mathematical Statistics* 13, 14-33.
- Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* 38, 963-974.
- Lawless, J.F. (1982). *Statistical models for lifetime data*. New York: Wiley.
- Lee, Y.K. (1974). A note on Rao's reduction of Pothoff and Roy's generalized linear model. *Biometrika* 61, 349-351.
- Lehmann, E.L. (1917). *Nonparametrics: statistical methods based on ranks*. Holden Day, San Francisco.
- Machen, D. (1975). On a design problem in growth studies. *Biometrika* 31, 749-753.
- Morrison, D.F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.
- Peto, R. (1972). Contribution to the discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society B*, 34, 205-207.

- Pothoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51, 313-326.
- Prentice, R.L. (1978). Linear rank tests with right-censored data. *Biometrika* 65, 167-179.
- Prentice, R.L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* 35, 861-886.
- Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* 52, 447-458.
- Rao, C.R. (1966). Covariance adjustment and related problems in multivariate analysis. In: Krishnaiah, P.R. ed.; *Multivariate Analysis*. New York: Academic Press, 87-103.
- Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2, 110-114.
- Timm, N.H. (1980). Multivariate analysis of repeated measurements. In: Krishnaiah, P.R. ed.; *Analysis of Variance*. New York: North Holland, 41-87.