WRITTEN EXAMINATIONS FOR THE MASTER OF SCIENCE DEGREE

in the

Department of Biostatistics
School of Public Health
University of North Carolina at Chapel Hill

assembled and edited by
DANA QUADE
Institute of Statistics Mimeo Series No. 1465

First edition: September 1984
Second edition: January 1988

WRITTEN EXAMINATIONS

for the degree

MASTER OF SCIENCE

in the

DEPARTMENT OF BIOSTATISTICS

School of Public Health

University of North Carolina at Chapel Hill

assembled and edited by

DANA QUADE

First edition: September 1984

Second edition: January 1988

TABLE OF CONTENTS

				Page
Introduction				1
12	April	1980		4
11-12	April	1981		12
3-4	April	1982		19
29	May	1982	("Special offer", Part II only)	28
15-16	January	1983		29
9-10	April	1984	("Special")	40
21-22	January	1984		49
15	April	1984	(Part II only)	58
26-27	January	1985		66
13-14	April	1985	("Special")	76
18-19	January	1986		84
12-13	April	1986	("Special")	92
14-15	March	1987		98
7-8	May	1987		108

INTRODUCTION

This publication contains the written examinations which the Department of Biostatistics has set for candidates for the degree Master of Science (MS), beginning in 1980. Prior to that time, MS candidates took the "Basic Doctoral Written Examination" as their Master's Written Examination, although the standard for acceptable performance was set lower for them than for doctoral candidates. (Copies of the Basic Examination are available in the Institute of Statistics Mimeo Series as Issues #1343: Closed-book Parts and #1344: Take-home Parts.) The Departmental Examinations Committee prepares and conducts all Department-wide written examinations, and arranges for their grading.

During the years 1980 through 1982 the MS examination consisted of two Parts, Part I being closed-book, and Part II open-book. Each Part consisted of 5 questions, of which #4 and #5 were considered more difficult than the others; the MS candidate was expected to answer 3 of the 5 questions, within a period of 3 hours. (MPH and MSPH candidates took the same examination, but under more lenient rules; see Mimeo Series #1329.)

Beginning in 1983 the MS examination has been entirely separate. There are two Parts: Part I (Theory) is closed-book, and Part II (Applications) is open- book. In each part there are 4 questions, of which the candidate is to answer 3. The time limit was originally set at 3 hours for each Part; but, starting in 1985, 4 hours are now allowed for Part II. This examination is given each Spring. Special re-examinations for the MS have been combined with the MPH examination later in the year, however, and these have been conducted under different rules; see the copies of the examinations which follow.

A team of two graders is appointed for each question. Where possible, all graders are members of the Department of Biostatistics and of the Graduate Faculty, and no individual serves on more than one team for the same examination (the two Parts counting separately in this context). The members of each grading team prepare for their question an "official answer" covering at least the key points. They agree beforehand on the maximum score possible for each component, the total for any question being 25. The papers are coded so that the graders are unaware of the candidates' identities, and each candidate's answer is marked independently by each of the two graders. The score awarded reflects the effective proportion correctly answered of the question. The two graders then meet together and attempt to clear up any major discrepancies between their scores. Their joint report may include comments on serious shortcomings in any candidate's answer.

On the basis of a condidate's total score on a paper, the Examinations Committee recommends to the faculty whether the candidate is to be passed, failed, or passed conditionally. In the last case, the condition is specified, together with a time limit. All final decisions are by vote of the faculty. Examination papers are not identified as to candidate until after the verdicts of PASS and FAIL have been rendered. Once the decisions have been made, advisors are free to tell their students unofficially; the official notification, however, is by letter from the Chairman of the Examinations Committee. Actual scores are never released, but the "official answers" are made public, and candidates who are not passed unconditionally are permitted to see the graders' comments on their papers. A candidate whose performance is not of the standard required may be reexamined at

the next regularly scheduled examination, or at an earlier date set by the Examinations Committee. One reexamination is permitted automatically.

Candidates whose native language is not English are not to be allowed extra time on Department-wide (not individual course) examinations. This condition may be waived for individual candidates at the discretion of the Department Chairman upon petition by the candidate at least one week prior to the examination.

NOTE. Most of what follows reproduces the examinations exactly as they were originally set; however, minor editorial changes and corrections have been made, particularly in order to save space.

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(April 12, 1980)

Question 1 The following questions apply to stratified sampling:

a) Briefly describe stratified (simple) random sampling.

b) What makes stratified random sampling a type of probability sampling?

c) Briefly discuss the set of guidelines which you would use in forming strata in this type of design.

d) Briefly describe the relationship between an epsem (or self-weighting) design and proportionate stratified sampling.

e) Noting that the estimated variance of $\bar{y}_o = (1/n)\sum_j^n y_j$ from a simple random sample of size n from a population of size N is

$$\mathrm{var}(\bar{y}_o) = \left(\frac{1-f}{n}\right)s^2$$

where

$$s^2 = \frac{\sum_j^n (y_j - \bar{y}_o)^2}{n-1} \quad \text{and} \quad f = (n/N),$$

derive an estimator for the variance of the estimated mean from a stratified random sample,

$$\bar{y}_{wo} = \sum_h^H W_h \bar{y}_{ho}$$

where $W_h = (N_h/N)$, N_h is the number of elements in the h-th stratum of the population, $N = \sum_h^H N_h$, H is the total number of strata,

$$\bar{y}_{ho} = \frac{\sum_j^{n_h} y_{nj}}{n_h}$$

and n_h is the number of elements in the sample from the h-th stratum.

Question 2 Consider the regression model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n.$$

Suppose that you want to test

$$H_0: \beta_2 = 0 \quad \text{vs.} \quad H_1: \beta_2 > 0.$$

a) State the usual assumptions you need to make for this purpose.

b) Derive the expression for the mean square due to error.

c) Write down the formula for the test statistics.

d) Suppose that you want to test also

$$H_0^*: \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_1^*: (\beta_1, \beta_2) \neq \underset{\sim}{0}.$$

e) What are the critical regions for the testing problems in (c) and (d)?

Question 3 A large machine consists of 50 components. Past experience has shown that a particular component will fail during an 8-hour shift with probability .1 . The equipment will work if no more than one component fails during an 8-hour shift.

a) Calculate the probability that the machine will work throughout an entire 8-hour shift, assuming the binomial distribution is applicable. Carefully define any notation and state the assumptions required for the valid application of the binomial procedure.

b) State the general situation and assumptions which are required for the Poisson model to be applicable to this situation. Give the Poisson density function.

c) The Poisson distribution may be used to approximate the binomial distribution when n is "large" and p is "small". Use the Poisson distribution to find an approximation to the probability computed in part (a).

d) Calculate an approximation to the probability obtained in part (a) by using the normal approximation to the binomial distribution. State the assumptions under which the approximation is reasonable.

e) Compare the accuracy and usefulness of these three calculations.

Question 4 Suppose that X has a chi-square distribution

$$f_X(x) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} x^{\nu/2 - 1} e^{-x/2}, \quad x > 0, \quad \nu > 0,$$

and Y has a beta distribution

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad \alpha > 0, \quad \beta > 0$$

$$0 < y < 1.$$

a) Obtain formulae for $E(X^r)$ and $E(Y^r)$, where r is a positive integer.

b) If $r < [\frac{\nu}{2}]$ * show that

$$E(X^{-r}) = [(\nu-2)(\nu-4)\cdots(\nu-2r)]^{-1} .$$

c) Find $E(Y^{-r})$. What condition must r satisfy?

d) Assume that X and Y are independent, and let $Z = X/Y$. Find $E(Z)$.

Question 5 Suppose that lifetime (T) of a certain mechanical device has a Weibull distribution with PDF

$$f_T(t) = \frac{c}{\theta^c} t^{c-1} \exp[-(\frac{t}{\theta})^c] \quad \theta > 0, \quad c > 0, \quad t > 0$$

a) Obtain the formula for cumulative distribution function, $F_T(t)$, and evaluate the expected proportion of failures

(i) before time τ;

(ii) after time τ.

b) Suppose that N items were put on test at $t = 0$, and n were observed to fail before time τ. Suppose that exact failure times were recorded; let t_i denote the failure time of the i-th item among those items which failed $(i = 1,\ldots,n)$. Construct the likelihood function.

* $\lfloor s \rfloor$ denotes the largest integer $\leq s$.

EDITORIAL NOTE: Two tables were appended to this examination:
a) Standard normal distribution function $\Phi(x)$ for $x = -3(.01)3$
b) Natural logarithms $\ln(x)$ for $x = 1(.01)10$

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(April 12, 1980)

M.P.H. and M.S.P.H. students are to answer *any two* questions during the two-hour period (1 pm - 3 pm). M.S. students are to answer *three* questions of which not more than 2 should be from Section A - time period 1 pm - 4 pm

You are required to answer *only what is asked* in the questions and *not all you know* about the topics.

Question 1 A survey of 320 families, each with 5 children, revealed the following distribution:

No. of girls	0	1	2	3	4	5	Total
No. of families	18	56	110	88	40	8	320

a) Is the result consistent with the hypothesis that male and female births are equally probable?

Test this hypothesis at the significance level $\alpha = 0.05$, $\alpha = 0.01$

b) What is the maximum likelihood estimate of the probability of a female birth?

Question 2 A. Briefly describe or explain the following terms:

(a) OS
(b) TSO
(c) Track (on magnetic disk)
(d) Block
(e) Logical Record
(f) byte
(g) JCL

B. Compare and contrast:

(a) OS dataset
(b) SAS database
(c) SAS dataset

(That is, demonstrate that you know what each of these terms means and the differences between them.)

C. The printout in Figure 1 was produced by the PROC PRINT statement in line 190 of the SAS program shown in Figure 2. Show what will be printed as a result of

(a) the PROC PRINT statement in line 230 in Figure 2. (Be careful to write out the entire output to be produced by SAS *except* titles and page headings.)

(b) Show what will be printed as a result of the PROC MEANS (Figure 2, line 250) and related statements.

(c) Show what will be printed as a result of the PROC PRINT statement in line 440 of Figure 2.

(d) Show what will be printed as a result of the PROC PRINT statement in line 530 of Figure 2.

Figure 1. Printout Produced by the PROC PRINT Statement in Line 190 of Figure 2.

S T A T I S T I C A L A N A L Y S I S S Y S T E M

OBS	NAME	SEX	AGE	HEIGHT	WEIGHT
1	ALFRED	M	14	69	112
2	ALICE	F	13	56	84
3	BERNADETTE	F	13	65	98
4	BARBARA	F	14	62	102
5	JAMES	M	12	57	83

Figure 2. SAS Program for Problem 2

```
00050 //     EXEC SAS
00060 //*PW=EXAM
00070 //REX DD *
00080 M 14 69 112 ALFRED
00090 F 13 56 84 ALICE
00100 F 13 65 98 BERNADETTE
00110 F 14 62 102 BARBARA
00120 M 12 57 83 JAMES
00130 //SYSIN DD *
00140 DATA STUDENT1;
00150    INFILE REX;
00160    LENGTH NAME $ 20 SEX $ 1;
00170    INPUT SEX AGE HEIGHT WEIGHT NAME;
00180
00190    PROC PRINT;  * THIS PRODUCES FIGURE 2;
00200
00210 PROC SORT; BY SEX AGE HEIGHT;
00220
00230 PROC PRINT;  * PROBLEM 3(A);
00240
00250 PROC MEANS MEAN N;  * PROBLEM 3(B);
00260    BY SEX;
00270    VAR AGE HEIGHT;
00280
00290 PROC SORT; BY NAME SEX AGE;
00300
00310 DATA CHANGES;
00320    LENGTH NAME $ 20 SEX $ 1;
00330    MISSING _;
00340    INPUT SEX AGE HEIGHT WEIGHT NAME;
00350  CARDS;
00360    F 13 50 . ALICE
00370    M 12 . 89 JAMES
00380    M 14 _ .  ALFRED
00390    F 14 -1 . BARBARA
00400    F 16 62 102 BARBARA
00410    F 12 59 84 JANE
00420  ; * END OF CHANGE DATA;
00430
00440 PROC PRINT;     * PROBLEM 3(C);
00450
00460 PROC SORT; BY NAME SEX AGE;
00470
00480 DATA STUDENT2;
00490    UPDATE STUDENT1 CHANGES;
00500    BY NAME SEX AGE;
00510    IF HEIGHT > 0;
00520
00530 PROC PRINT;  * PROBLEM 3(D);
```

Question 3 A sample survey of 800 adults is conducted in a large city in order to determine citizen attitudes toward national health insurance. The sampling frame for the survey is a list of adult taxpayers from which four strata are formed corresponding to each of four major sections of the city. The design calls for selecting a simple random sample within each stratum. General data for the survey sample as well as results for an opinion question on national health insurance are as follows:

	Stratum (h)				
	(Inner City)	(Blue Collar)	(Middle Class Suburbs)	(Affluent Suburbs)	TOTAL
Total number of adults (N_h)	20,000	50,000	20,000	10,000	100,000
Number of adults interviewed in the sample (n_h)	200	200	200	200	800
Number of interviewed sample adults who are in favor of national health insurance (x_h)	190	180	40	30	440

a) Calculate a stratified estimate of the proportion (P) of adults in the city who favor national health insurance.

b) Calculate an estimate of the variance of the estimate produced in Part (a).

c) A colleague argues that since a "random sample" of adults has been chosen and since simple random sampling was used, the estimate and its variance can be calculated as if a simple sample of $n = 800$ adults had been selected. Calculate the estimate and its variance as the colleagues suggests.

d) Comment briefly on the difference between your analysis and your colleague's analysis.

e) If you decide to use Neyman allocation for a similar survey on national health insurance in the future, how would you then allocate a sample of $n = 800$ adults to the same four strata?

Question 4 Read the attached article by Topoff and Mirenda from a recent issue of *Science*. Note that from the data of Table 1 the authors were led to calculate a chi-squared statistic. Discuss the statistical appropriateness and the numerical accuracy of their analysis.

EDITORIAL NOTE: The article referred to, including Table 1, was reproduced and appended to this examination. It originally appeared in *Science 207:* 1099-1100 (7 March 1980).

Question 5 Ten observations on three independent variables (X_1, X_2, X_3) and one dependent variable were obtained as follows:

X_1	X_2	X_3	Y
2.7	5.5	6.0	2.3
2.2	6.2	6.8	1.8
2.7	5.5	6.0	2.7
3.0	6.2	7.3	2.7
3.2	7.9	7.1	2.8
2.2	6.2	6.8	2.2
2.7	5.5	6.0	2.4
4.0	9.0	7.8	1.6
3.0	6.2	7.3	2.9
4.0	9.0	7.8	2.2

It is postulated that the model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1^2 + \beta_{32} X_3^2 + \beta_{33} X_3^3$$

is reasonable.

(i) Inspect the model and the data and comment on the appropriateness of the model.

(ii) Determine an estimate of the variance of the random error, σ^2, for these data.

(iii) By plotting the data or otherwise, select a model with few parameters that appears appropriate and fit that model. Comment on the fit.

(iv) Does R^2 tell you the same thing as a direct test of lack of fit based on an estimate of pure error?

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(April 11, 1981)

Question 1

a) Define probability sampling from finite populations.

b) Briefly discuss the advantages of probability samples over nonprobability samples.

c) The recreation department in a large city hires you to design a cluster sample of families for an interview survey. The most important survey objective is to produce a 95% confidence interval for the proportion (P) of families who are aware of city recreation programs. The expected "half-width" of the confidence interval for the estimated proportion (p) will be about*

$$d = 1.96\sqrt{Var(p)}$$

You decide to select a simple random sample of blocks. Your past experience tells you that the design effect will be about 4. You also know that P is about 0.3. Ignoring the finite population correction, how large a sample of families would you recommend so that d = 0.05?

Question 2 Let X and Y have the joint density function

$$f_{X,Y}(x,y) = 2(1-\theta)^{-2},\ 0 < \theta < y < x < 1.$$

a) Find the marginal distribution of X.

b) Find the conditional distribution of Y given that X = x.

c) The quantity $E(Y|X = x)$ is often called the <u>regression function</u> of Y on x. Find $E(Y|X = x)$.

d) For n pairs of data points (x_i, Y_i), i = 1, . . ., n, derive the formula for the <u>least squares</u> estimator of θ using your answer to part (c).

---

*Recall that a 95% confidence interval for the estimated proportion (p) produced in your analysis will be of the form:

Lower Limit: $p - 1.96\sqrt{Var(p)}$

Upper Limit: $p + 1.96\sqrt{Var(p)}$

Question 3 Suppose that 60% of a particular breed of mice exhibit aggressive behavior when injected with a given dose of a stimulant. An experimenter will apply the stimulant to 3 mice one after another and will observe the presence or absence of aggressive behavior in each case.

a) List the sample space for the experiment. (Use A to denote aggressive and N to denote nonaggressive).

b) Assuming that the behaviors of different mice are independent, determine the probability of each elementary outcome.

c) Find the probability that
(i) two or more mice will be aggressive,
(ii) exactly two mice will be aggressive, and
(iii) the first mouse will be nonaggressive while the other two will be aggressive.

d) What can you say about the exact distribution of the signed rank statistic when H_o holds and n is small? What approximation would you recommend when n is large?

Question 4 Let $f_X(x) = \theta^{-1}e^{-x/\theta}$, $x > 0$, $\theta > 0$ be the pdf of an exponential distribution.

a) Show that, in fact, X is distributed as $(\theta/2)\chi_2^2$ where χ_2^2 has the chi square distribution with 2 degrees of freedom .

b) Based on a random sample $X_1, \ldots, X_n$ of size n, find the maximum likelihood estimator of θ.

c) Construct the likelihood ratio test for testing the null hypothesis H_0: $\theta=\theta_o$ against the alternative $\theta<\theta_o$.

d) Derive the power function of the test.

Question 5 Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables with an unknown (but continuous) distribution function F and θ be the median of F. Suppose that one wants to test for

$$H_0: \theta = \theta_o \text{ (Specified) against } H_1: \theta > \theta_o.$$

a) Write down the expression for the sign test statsitic for this testing problem and its exact distribution under H_0. What approximation to this distribution would you recommend when n is large?

b) What additional assumption do you need to make to use the Wilcoxon signed rank statistic for this testing problem?

c) Compute the first two moments of the Wilcoxon signed-rank statistic (under H_o).

PART II

(April 12, 1981)

Question 1 The mean drying time of a brand of spray paint is known to be 90 seconds. The research division of the company that produces this paint contemplates that adding a new chemical ingredient to the paint will accelerate the drying process. To investigate this conjuncture, the paint with the chemical additions is sprayed on 15 surfaces and the drying times are recorded. The mean and standard deviations computed from these measurements are 86 seconds and 5.6 seconds respectively.

5 points (a) Do these data provide strong evidence that the mean drying time is reduced by the addition of the new chemical?

6 points (b) Construct a 98% confidence interval for the mean drying time of the paint with the chemical additive.

9 points (c) Suppose that the actual standard deviation for the drying time does not change with the addition of the new chemical and is known to be equal to 6 seconds. Given this additional information, what would be your conclusions in (a) and (b)?

5 points (d) Suppose that it is also conjectured that the standard deviation of the drying time decreases with the addition of the new chemical. Do these data provide a strong evidence for that?

Question 2 The state welfare agent is in the process of sampling unemployment data in his state. The state is divided into 4 regions each with approximately the same population. Each region is in turn divided into 750 equal-sized sampling units. From each region five sampling units are selected and sampled intensively. The percentage unemployment for one such test is given below.

Region A:	4.2	4.4	4.5	5.0	5.1
Region B:	3.7	3.9	4.1	4.4	4.5
Region C:	4.8	5.0	5.1	5.2	5.2
Region D:	3.1	3.5	3.6	3.7	3.9

(a) Is this a simple random sampling? Why or why not?

(b) Calculate the mean unemployment rate for each region and use them to estimate the mean rate for the entire sample of 20 obervations.

(c) Compute the mean using the entire 20 observations. Does it differ from the answer in part (b)? Why or why not?

(d) What procedure would be necessary if the regions and sampling units were of different sizes in terms of population?

Question 3

A) Briefly describe the purpose(s) of:

(a) JCL

(b) DATA step of SAS

(c) PROC step of SAS

B) Define, and describe the relationships between:

OS file (dataset)

SAS database

SAS dataset

(Examples of corresponding JCL and SAS code may be useful.)

C) In less than 1 page, outline the basic steps of the process required to create a SAS dataset that is ready for statistical analysis. Suppose the input dataset is a raw, unchecked dataset stored on disk. The SAS dataset is also to be stored on disk.

D) Write out the job or jobs, including JCL and SAS code, needed to do the following on a dataset with the format given in Figure 1. Use UNC.B.E.99U as the account number and MEXAM as the password.

(a) Create a SAS dataset, stored on on-line disk, called VEHACC that includes all the variables, but just the reportable cases. Use the variable names given in capitals in the format. Label Height and Weight.

(b) Print 10 observations.

(c) Create a variable called AGE_GP with the following codes:

AGE_GP values	AGE values
1	0 thru 10
2	11 thru 24
3	25 thru 54
4	55 and over
.	Unknown

(d) Plot Height vs. Weight for males and females separately.

(e) Create cross-tabulation tables of Sex by Restraint, Sex by Injury, and Injury by Restraint.

Figure 1

Format for a vehicle-oriented accident file

DSN=UNC.B.E999U.MASTERS.VEH.RAW

There are approximately 1000 records

Column	Information
1	Accident Reporting Type - ATYPE 1 Non-reportable 2 On private property 3 Reportable
2-7	Accident Case Number - ID
8	Injury Class - INJ 1 Not injured 2 Class C injury 3 Class B injury 4 Class A injury 5 Killed 6 Not stated
9	Restraint Used - BELT 1 No belt 2 Lap and shoulder belt 3 Child restraint 4 Not stated
10	Race - RACE 1 White 2 Black 3 Indian 4 Other 5 Not stated
11	Sex - SEX 1 Male 2 Female 3 Not stated
12	Age - AGE 01-97 Actual age 98 Older than 97 99 Not stated
13-14	Height - HT 01-98 Actual height in inches 99 Not stated
15-17	Weight - WT 001-998 Actual weight in pounds 999 Not stated

Question 4 Suppose we have conducted an experiment to estimate the weight gain for a sample of 14 dairy cows as a result of a one week exposure to a feed additive. One statistical question of interest is to test whether the weight gain is zero. Another major objective is to estimate the weight gain.

The 14 dairy cows comprised three different breeds; there were 7 Holsteins, 5 Jerseys, and 2 Guernseys. The data collected were as follows:

	Weight Gain in lbs.
Holsteins	-7, 5, -1, 3, 1, 6, 0
Jerseys	2, -2, 1, 7, 2
Guernseys	9, 2

1. Suppose we were interested in testing H_0: mean weight gain = 0, and we had in mind a target population of dairy cows in which Holsteins, Jerseys, and Guernseys were in the ratio 7:5:2. Specify the test you would carry out. Compute the test and state the significance level. Provide a corresponding estimate of mean weight gain.
2. Suppose we were interested in testing H_0: mean weight gain = 0, and we were primarily interested in a target population of dairy cows in which Holsteins, Jerseys, and Guernseys were in equal proportions. Specify the test you would carry out. Compute the test. Provide a corresponding estimate of mean weight gain.
3. Are tests in questions 1 and 2 the same? Comment.
4. Suppose now that we were interested in testing H_0: mean weight = 0 and we had a target population in which Holsteins, Jerseys, and Guernseys were in the ratio $W_H:W_J:W_G$. Specify an appropriate test.

Question 5 In measuring the various constituents of cow's milk, it is of interest to determine how protein (Y) is related to fat (x_1) and solid-nonfat (x_2). Samples of 10 cows were taken and the following data were obtained:

Observations	Protein Y	Fat x_1	Solid Non-Fat x_2
1	3.75	4.74	9.50
2	3.19	3.66	8.56
3	2.99	4.27	8.54
4	3.46	4.03	8.62
5	3.27	3.51	9.35
6	3.27	3.97	8.39
7	2.78	3.23	7.87
8	3.59	3.79	9.33
9	3.16	3.36	8.86
10	3.65	3.64	9.21

Assume the linear regression of Y on x_1, x_2 and

(i) Find the least squares estimates of the (partial) regression coefficients.

(ii) We are interested in testing the null hypothesis that both these partial regression coefficients are equal to 0. Test this hypothesis at the significance level $\alpha = 0.05$. State clearly the assumptions you need to make in this context.

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(April 3, 1982)

Question 1.

You are called upon to assist the health department in a large city with the design of a local household survey. The survey's principal objective will be to estimate the proportion of households in which the person usually responsible for preparing the meals is aware of the importance of a balanced dietary intake by members of the household. Between 20 and 40 percent of the local households are thought to be aware of this. The health department recognizes the importance of a high response rate, but has only a modest amount of money to do the survey. Moreover, the survey will probably have to be conducted by staff of the health department who collectively have little survey experience.

a. *Briefly* discuss the relative merits of the three methods of data collection being considered for the survey: self-administered questionnaire by mail, telephone interview, and personal interview.

b. If either telephone or personal interviewing is the selected method, two-stage cluster sampling will be used to select the sample of households. The design effect in either case is expected to be about 1.5. Briefly describe what a "design effect" is and what things contribute to its size.

c. In the event that the two-stage design is used, determine the number of completed household interviews which would be required to yield a coefficient of variation of 10 percent. You can ignor the finite population correction.

[Hint: Recall that the coefficient of variation for the estimator (p) of the true population (P) is

$$CV(p) = \frac{[Var(p)]^{\frac{1}{2}}}{P}$$]

Question 2.

Let U be a random variable with p.d.f.

$$f(u) = 1, \quad 0 < u < 1 \quad .$$

(a) Find the p.d.f. of $X = -\lambda \log(1-U)$.

(b) Find $E(X)$.

(c) Let Z be a random variable (independent of U) with p.d.f.

$$g(z) = (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2} z^2), \quad -\infty < z < \infty \ .$$

Let $Y = e^{-Z}$ and $W = -Y \log(1-U)$.

Find $E(W)$.

Question 3.

An executive is willing to hire a secretary who has applied for a position unless a significnace test indicates that she averages more than one error per typed page. A random sample of five pages is selected from some typed material by this secretary and the errors per page are: 3, 4, 3, 1, 2.

(a) Assuming that the numbers of errors, λ, say, per page has Poisson distribution, what decision will be made? Use significance level $\alpha \leq 0.05$.

(b) Calculate the power when the average number of errors per page, λ, is equal to 2.

(c) Suppose that the executive looks at a random sample of 225 pages of the secretary's work and finds 252 errors. What would be his decision, using $\alpha \doteqdot 0.05$?

Individual terms, $e^{-m} m^i/i!$, of the Poisson distribution

	m			
i	1	2	5	10
0	.36788	.13534	.00674	.00005
1	.36788	.27067	.03369	.00045
2	.18394	.27067	.08422	.00227
3	.06131	.18045	.14037	.00757
4	.01533	.09022	.17547	.01892
5	.00307	.03609	.17547	.03783
6	.00051	.01203	.14623	.06306
7	-	.00344	.10445	.09008
8	-	.00086	.06528	.11260
9	-	-	.03627	.12511
10	-	-	.01813	.12511
11	-	-	.00824	.11374
12	-	-	.00343	.09478
13	-	-	.00132	.07291
14	-	-	.00047	.05208

Normal percentile point

$\tau_{.95} = 1.645$

$\tau_{.90} = 1.282$

$\tau_{.975} = 1.960$

Question 4.

Suppose that a system has two components whose *life times* (X and Y, say) are independent and each has the same exponential distribution with mean $\theta(> 0)$. The system fails as soon as at least one of its components does so. Let Z be the life-time of the system.

(a) What is the probability density function of Z?

(b) For $n(\geq 1)$ systems of the same type, let $Z_1,\ldots,Z_n$ be the respective life times. Obtain the maximum likelihood estimator of θ (say, $\hat{\theta}_n$) based on $Z_1,\ldots,Z_n$.

(c) Obtain $E(\hat{\theta}_n)$ and $Var(\hat{\theta}_n)$.

Question 5.

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be n independent bivariate observations from a continuous bivariate distribution $F(x,y)$, $-\infty < x,y < \infty$. Let H_0 be the null hypothesis that X and Y are independent.

(a) Define $\tau = 2\Pr\{(X_1-X_2)(Y_1-Y_2) > 0\} - 1$ and show that under H_0, $\tau = 0$.

(b) Obtain the symmetric and unbiased estimator (t_n) of τ based on the n observations and deduce the expressions for $E(t_n|H_0)$ and $V(t_n|H_0)$.

(c) What can you say about the large sample distribution of $n^{\frac{1}{2}}t_n$, when H_0 holds?

(d) What modifications to t_n would you suggest to accommodate possible ties among the X's and/or the Y's?

PART II

(April 4, 1982)

Q.1. An investigator is planning a study to evaluate a new medication for the treatment of hypertension. She knows from past experience that for patients with hypertension the mean diastolic blood pressure is 105 mm., the standard deviation is 15 mm. and the correlation between two measurements is 0.7.

(a) Find the sample size needed if she uses each patient as his own control. Assume $\alpha = 0.05$, $\beta = 0.1$, a one-sided test and that she wants to detect a change in blood pressure of 10 mm.

(b) Suppose she takes a group of patients and randomly divides them into two groups. She will give one group the new drug and the other group will get no treatment. If she will compare the change in the blood pressure in the treated group to that in the control group, what sample size does she need? Use the same assumptions as above.

(c) Discuss the relative merits of the two designs.

(d) Suppose the design in part (b) is chosen and that a total of 10 patients will be used. Use the attached table of random numbers to prepare a randomization schedule such that 5 patients will be assigned to treatment and 5 to control. Please give details about how the table is used so that the grader can reconstruct your schedule.

EDITORIAL NOTE: An attached table presented 3000 random digits in 60 rows of 50 digits each.

Q.2. An experiment was conducted to determine whether selenium supplementation is associated with reduced incidence of benign ovarian tumors in pregnant cows. One treatment group and one control group, of approximately equal sizes ($N \doteq 25$ in each) were used. Each cow in the treatment group received the same amount of selenium, injected once, a fixed number of weeks before the end of pregnancy.

To verify that the treatment raised the blood levels of two important proteins throughout pregnancy, for each cow blood samples were taken before injection and after the end of pregnancy. The concentrations of the proteins were determined at each of these two times for each cow. The important questions of interest here are whether blood levels were similar in the two treatment groups before treatment, whether these blood levels changed between treatment and the end of pregnancy within each treatment group, and whether the change was greater in the treated than in the control group if the latter also had a change. (If the treatment is effective, blood levels should increase.)

For each protein, the investigators determined whether differences existed by the use of two-way ANOVA (with factors treatment, and time when blood drawn), followed by the use of Tukey's multiple comparison method with $P = .05$.

(a) Evaluate the method of analysis. Are the required assumptions met? Does the analysis answer the questions of interest?

(b) If you find the current analysis inappropriate, propose a better one, showing how to answer the primary questions with level $P = .05$. If you find the current analysis appropriate, discuss the use of Tukey's multiple comparison test vs. some other method to answer the primary questions.

Q.3.

A. Briefly describe the purpose(s) of:

(1) JCL

(2) DATA step of SAS

(3) PROC step of SAS

B. Define, and describe the relationships between:

OS file (dataset)

SAS database

SAS dataset

(Examples of corresponding JCL and SAS code may be useful.)

C. List the major components of a large modern computer (CPU, etc.), and briefly describe the functions of each component and describe the relationship among them. A simple diagram may help you in organizing your answer.

D. List each type of JCL statement and briefly describe the function of each. Write a valid job (or jobs) including at least one example of each type of statement.

Q.4. A dentist who was responsible for dental care of cerebral palsied children in a state institution wanted to determine whether he should recommend that electric toothbrushes be purchased for routine use by the patients. He wanted to be as objective as possible in arriving at a decision, and decided to consult with a statistician about designing an experiment to determine whether short term improvement in oral hygiene could be demonstrated. In answering the major question, "Should the purchase of electric toothbrushes be recommended for this institution?", there are other considerations over and above any real improvement in oral hygiene of patients which should be taken into account but these were ignored in designing the study.

Study Design and Conduct of Trial

It was decided that the study should be designed to determine whether brushing with electric toothbrushes resulted in "cleaner teeth" than brushing with regular toothbrushes during a two week period. First, a search of the literature for measures of tooth cleanliness resulted in a decision to use the debris index, which is an average of debris scores for six teeth,* as the response variable, i.e., the variable which is to be altered (hopefully) by "treatment". Next, factors which could (potentially) influence results were listed:

1. Age
2. Race
3. Sex
4. Degree of ability to care for teeth (brushes own teeth or brushed by nurse)
5. Initial level of cleanliness (debris index)
6. Placebo effects:
   (a) Attitudes and actions of children and nurses
   (b) Attitude and actions of examining dentist

Because only 35 children were available for the initial examination it was impractical to stratify (or control) on all of these variables. However, randomization, a way of balancing the effect of variables which cannot be controlled, was used.

The study was carried out as follows:

1. Each child was examined by the dentist and a pre-trial debris index was determined using the following 3x5 form for recording.

*Greene, J.C. and Vermillion, J.R. "Oral Hygiene Index - A Method for Classifying Oral Hygiene Status", JADA, b1:172-179, (Aug. 1960).

Name_____ No._____

Age_____ Sex_____ Race_____

Comments:

	Right	Ant	Left	Total
Max (B)				
Mand (L)				
Total				

Debris Index_____

2. The children were stratified by sex(ward), and degree of disability, i.e., divided into four groups:
   (a) Male - brushes own teeth
   (b) Male - assisted by nurse
   (c) Female - brushes own teeth
   (d) Female - assisted by nurse

3. Within each group children were randomly assigned to one of two brushing groups:
   (a) Electric toothbrush
   (b) Regular tooth care

   The assignments were not disclosed to the dentist, i.e., he was "blind" as to type of care each child received.

4. A list of children assigned to the two groups was posted in each ward and the nurses supervised (and assisted where necessary) to see that assignments were followed.

5. At the end of the two week trial another examination was made by the dentist, who followed the same procedure as in the pre-trial examination to determine a debris index for each child. Results were recorded on another 3x5 card without reference to results of the original examination.

6. The results were matched with those from the first examination. Actual results are shown on the following page.

PROBLEM

(a) Test the statistical significance of the decline in debris index observed in each group, and with electric toothbrushing as compared with regular.

(b) Write a brief report, aimed at the dentist and the director of the state institution, describing the results and their analysis.

Results of Two Week Trial of Regular and Electric Tooth Brushing, in Cerebral Palsy Hospital Patients

Regular Tooth Brush

Child Number	Debris Index		
	Before	After	Difference
2	2.0	1.0	-1.0
9	1.5	0.8	-0.7
12	1.0	1.5	+0.5
16	1.5	Disch.	-
18	0.5	1.0	+0.5
28	2.5	2.2	-0.3
31	2.2	1.2	-1.0
34	1.7	1.0	-0.7
4	2.5	2.2	-0.3
14	0.7	0.5	-0.2
39	2.0	2.0	-0
20	1.8	0.8	-1.0
24	1.8	0.8	-1.0
26	1.5	1.8	+0.3
29	0.7	0.5	-0.2
32	1.0	0.7	-0.3
33	2.8	2.3	-0.5
Sum	27.7	20.3	-5.9
Sum of squares (uncorrected)	52.53	31.81	6.17
N	17	16	16

Electric Tooth Brush

Child Number	Debris Index		
	Before	After	Difference
1	1.0	1.3	+0.3
3	2.2	Disch.	-
11	0.8	Disch.	-
15	1.3	Disch.	-
22	1.5	1.2	-0.3
38	1.7	1.5	-0.2
25	2.5	1.0	-1.5
30	2.3	Disch.	-
21	2.0	0.8	-1.2
23*	2.2	0.8	-1.4
36	1.7	1.3	-0.4
5	0.9	Disch.	-
6	1.0	1.3	+0.3
7	2.0	1.5	-0.5
17	2.5	2.0	-0.5
19	2.2	0.7	-1.5
10	1.0	1.2	+0.2
35	2.9	Disch.	-
Sum	31.7	14.6	-6.7
Sum of squares (uncorrected)	62.89	19.22	8.91
N	18	12	12

*Child #23 was originally assigned to regular care group but transferred by nurses to electric toothbrush group.

Note: "Disch." refers to children who were discharged before post-trial examination.

Q.5. Consider the following data (n=5) for a dependent variable Y and a *carrier* X (i.e., independent variable)

i	X	Y
1	3	5
2	8	17
3	2	5
4	9	18
5	7	13

$\Sigma X = 29, \quad \Sigma Y = 58, \quad \Sigma XY = 414, \quad \Sigma X^2 = 207, \quad \Sigma Y^2 = 832$

(i) Fit the simple linear regression model

$$E(Y) = \beta_0 + \beta_1 X$$

(ii) Theory strongly suggests $\beta_0 = 0$. So fit the model $E(Y) = \beta_1 X$.

(iii) Are there any differences between models (i) and (ii) as regards numerical results? If so, please specify.

(iv) Suppose two further (X,Y) points, (0,1) and (0,½), were collected. What impact would these have on the estimate of β_1 for model (ii)?

(v) Would the impact of these two data points be the same for model (i)? Explain in one or two brief sentences.

(vi) For model (ii), define $h_i = \dfrac{X_i^2}{\sum_{k=1}^{n} X_k^2}$

(denominator is summing over the data points for X). Take h_i to be the *leverage* that the observation Y_i has on the predicted value $\hat{Y}$. This *leverage* is exerted through the spacing of the X values (i.e., the design), not through the actual observed value of Y_i. In general $\hat{Y}_i$ is a linear combination of Y's where the coefficients are h-like terms.

Write a few brief sentences interpreting the formula for h_i and comment if the definition of h_i is compatible with your answer to part (iv).

Note: If $h_i = 0$, the observed value Y_i has no influence whatever on the predicted value $\hat{Y}_i$. At the other extreme if $h_i = 1$, the predicted value $\hat{Y}_i$ will always be the observed value Y_i.

(vii) For model (i), define $h_i = \dfrac{1}{n} + \dfrac{(X_i - \bar{X})^2}{\sum_{k=1}^{n} (X_k - \bar{X})^2}$.

Again, interpret the formula for h_i in one or two sentences and comment whether the definition is compatible with your answer to part (v).

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

Special Offer: May 29, 1982

INSTRUCTIONS:

a) This is an *open book* examination.

b) M.P.H. and M.S.P.H. students are to answer *any two* questions during the two-hour period (1:30 pm - 3:30 pm). M.S. students are to answer *three* questions of which not more than 2 should be from Group A - time period 1:30 - 4:30 pm.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, *not* your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

f) You are required to answer *only what is asked* in the questions and *not all you know* about the topics.

EDITORIAL NOTE: This "special offer" was identical with the regular Part II given on 12 April 1981 (see pages 14-18).

BASIC M.S. WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(January 15, 1983: 9:30 AM to 12:30 PM)

Q.1 Suppose that a simple random sample of m households is chosen for a health survey in a city containing M households. Let y_α denote the number of persons in the α-th household who have been ill during the month prior to the survey, and let n denote the total number of persons in the α-th household. The estimator

$$r = \frac{y}{n} = \frac{\sum_{\alpha=1}^{m} y_\alpha}{\sum_{\alpha=1}^{m} n_\alpha} ,$$

is used to estimate the city's illness incidence rate,

$$R = \frac{Y}{N} = \frac{\sum_{\alpha=1}^{M} y_\alpha}{\sum_{\alpha=1}^{M} n_\alpha}$$

a. What is the selection probability for each person in the α-th household? Explain your answer.

b. Is this sampling design epsem or self-weighting with respect to individuals? Explain why or why not.

c. Show that the bias of r is

$$\text{Bias}(r) = -\rho_{rn}\{\text{Var}(r)\}^{\frac{1}{2}}\,\text{CV}(n) ,$$

where ρ_{rn} is the correlation between r and n, $\text{Var}(r)$ is the variance of r, and $\text{CV}(n)$ is the coefficient of variation of n.

Q.2 Researchers believe that the prevalence (Y) of byssinoses in workers in textile manufacturing plants is <u>linearly</u> related to the mean daily cotton dust level (X). Under the assumption that zero cotton dust level implies zero prevalence, the regression equation relating the mean of Y to X is

$$E(Y|X=x) = \beta x \quad .$$

a. Given the n pairs of data points (x_i, Y_i), $i = 1,2,\ldots,n$, chosen randomly from the conditional distribution

$$f_Y(y\ X=x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(y-\beta x)^2} \quad , \quad -\infty<y<+\infty \quad ,$$

show that the maximum likelihood estimator of β is

$$\hat{\beta} = \sum_{i=1}^{n} x_i Y_i \Big/ \sum_{i=1}^{n} x_i^2 \quad .$$

b. Show that the maximum likelihood estimator $\hat{\beta}$ of β found in part (a) is also the least squares estimator of β; in other words, show that $\hat{\beta}$ minimizes the function

$$\sum_{i=1}^{n} [Y_i - E(Y_i|X=x_i)]^2$$

with respect to β.

c. Show that $E(\hat{\beta}) = \beta$ given the conditional density in part (a).

d. Show that $V(\hat{\beta}) = \sigma^2 / \sum_{i=1}^{n} x_i^2$ given the conditional density in part (a).

e. Show that $\hat{\beta}$ is normally distributed given the conditional density in part (a).

Q.3 A scientist hypothesizes the following esoteric theory for cellular damage in humans due to radiation exposure. A particular cell in an individual is "hit" by radiation X times during the course of the individual's lifetime, with the number X of such hits having the Poisson distribution

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0,1,\ldots,\infty \quad .$$

For any such hit, there is a fixed probability p that some basic structural change will occur in the cell; also, the occurrences (or not) of structural changes for different hits are assumed to be mutually independent.

The random variable of interest is Y, the total number of structural changes that the particular cell undergoes during the course of the individual's lifetime.

a. What is $pr(Y=y|X=x_0)$? In other words, what is the <u>conditional distribution</u> of Y given that a particular cell experiences exactly x_0 hits?

b. Using the result in part (a), show that the unconditional distribution of Y, namely $pr(Y=y)$, is Poisson with mean λp.

c. The scientist further hypothesizes that a particular cell will become cancerous if it undergoes at least k structural changes. Given the result in part (b), what is the probability that a particular cell will become cancerous? (JUST SET UP YOUR ANSWER). [<u>NOTE</u>: You do <u>not</u> need to be able to work parts (a) and (b) in order to answer this question.]

Q.4 Let X be distributed as a Poisson variate with mean $p\lambda$, and independently, let Y be distributed as a Poisson variate with mean λ. The parameters p and λ are both positive.

a. Show that the distribution of X given $X + Y = m$ is the binomial one:

$$\binom{m}{x}\left(\frac{p}{1+p}\right)^x\left(\frac{1}{1+p}\right)^{m-x}, \quad 0 \le x \le m \quad .$$

b. Find the maximum likelihood estimator (MLE) of p from (a).

c. Find the asymptotic variance of the MLE, conditional on m.

d. Find the unconditional information in part (c).

BASIC M.S. WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(January 16, 1983: 2:00 to 5:00 PM)

Q.1 It has recently been discovered that certain blood chemistry measurements correlate highly with the clinical diagnosis of "depression". One of these, the DST test, is estimated to have 70% sensitivity and 96% specificity for depression; another, the TSH test, has low sensitivity, only 25%, but it seems to have 100% specificity, since no false positive has yet been documented.

1. Assuming that the two tests operate independently (which agrees with the limited evidence available), what sensitivity and specificity could you achieve by applying both to the same subject?

The above results were obtained in subjects who were not physically ill. There is interest in applying the tests to cervical cancer patients, in whom the clinical diagnosis of depression is difficult because depressive symptoms are easily confused with those of the cancer itself. You are called in to help design a study to find out whether the two tests have different sensitivity and specificity in such patients. You are told that between 2 and 3 new patients per week will be available, that about 25% to 35% of these patients will actually have depression by clinical criteria, and that the study can continue for about 6 to 8 months.

2. Discuss generally the points that you would stress as a statistician in talking with the principal investigator. And more specifically, perform some calculations to indicate what success the study is likely to have in meeting its objectives.

   Note: sensitivity = pr (+ test result/diseased)
   specificity = pr (- test result/not diseased)

Q.2 Twenty-six subjects with essential hypertension were classified as "low" or "normal" with respect to plasma renin activity (PRA) in 1974-75, using two different methods. The same subjects were re-examined and reclassified in 1982.

The data for this study are given in the accompanying table. A subject had "low" PRA by Method A (PRA-sodium index) in 1974-75 (in 1982) if A1 < .555 (A2 < .555); by Method B (PRA after furosemide) according as B1 or B2 < 1.75 (A and B do not measure PRA in the same units.)

1. Do the two methods of classification agree with each other?
2. Does the classification by any one method remain consistent over a long period of time (i.e., from 1974-75 to 1982)?
3. Do these data indicate any general decrease in PRA with increasing age (as has been found in other studies)?

   NOTE: For purposes of this exam, satisfactory answers can be given which require very little computation.

LISTING OF SELECTED VARIABLES

OBS	ID	RACE	SEX	DOB	A1	A2	B1	B2	A1-A2	B1-B2
1	1	2	2	28 JUL 28	0.478	0.138	0.60	0.49	0.340	0.11
2	2	2	1	07 MAR 33	0.581	0.721	1.00	2.24	-0.140	- 1.24
3	3	2	2	23 MAY 44	3.298	2.697	6.50	1.70	0.601	4.80
4	4	2	2	22 FEB 43	0.230	1.059	0.73	1.36	-0.829	- 0.63
5	5	2	2	05 JUN 41	0.370	0.769	3.70	2.37	-0.399	1.33
6	6	2	2	20 JUL 36	1.207	3.619	2.50	4.38	-2.412	- 1.88
7	7	2	2	12 JUL 37	1.145	0.152	1.40	1.60	0.993	- 0.20
8	8	2	2	24 OCT 24	1.233	0.339	0.42	0.43	0.894	- 0.01
9	9	2	2	24 DEC 27	0.058	0.065	0.40	0.05	-0.007	0.35
10	10	2	2	19 SEP 42	0.692	0.648	4.12	0.90	0.044	3.22
11	11	2	2	15 MAR 51	1.295	1.592	1.20	4.09	-0.297	- 2.89
12	12	2	1	28 AUG 39	1.361	0.208	0.50	0.95	1.153	- 0.45
13	13	2	2	02 SEP 22	0.744	0.144	0.53	1.16	0.600	- 0.63
14	14	2	2	23 MAR 38	0.986	1.053	4.40	1.04	-0.067	3.36
15	15	2	2	09 OCT 47	4.939	0.922	14.60	1.87	4.017	12.73
16	16	2	2	13 JUL 22	0.903	0.699	3.40	3.52	0.204	- 0.12
17	17	2	2	17 APR 34	1.926	2.723	2.00	2.36	-0.797	- 0.36
18	18	2	1	04 MAY 47	5.365	0.397	8.10	0.60	4.968	7.50
19	21	2	2	31 MAY 32	0.191	0.903	0.70	--	-0.712	--
20	22	2	2	26 DEC 33	2.284	1.231	3.80	3.43	1.053	0.37
21	23	2	1	27 JUL 27	0.043	0.344	1.00	4.20	-0.301	- 3.20
22	25	2	1	25 OCT 31	1.323	0.253	3.14	0.53	1.070	2.61
23	26	2	2	25 DEC 27	0.402	2.419	1.00	4.05	-2.017	- 3.05
24	27	1	1	15 AUG 27	0.927	2.341	4.00	5.50	-1.414	- 1.50
25	28	2	1	09 APR 34	1.229	2.119	6.00	2.40	-0.890	3.60
26	29	1	1	27 SEP 38	2.188	2.051	7.60	--	0.137	--

Notes: 1) RACE 1 = white, 2 = black
2) SEX 1 = male, 2 = female
3) DOB = date of birth

Q.3 In a paper by Chalmers *et al*. (NEJM, 1977) entitled "Evidence Favoring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction", it was suggested that the negative findings in five of six randomized control trials may have been a function of small sample sizes. When the data from all six studies were "pooled", the apparent reduction in mortality with anticoagulator was 4.2%, which was found to be statistically significant. As provided in this paper, the data for these six trials can be shown as follows:

TRIAL	ANTICOAGULATED			CONTROLS		
NUMBER	TOTAL	DIED	%	TOTAL	DIED	%
1	45	13	28.9	47	18	38.3
2	77	12	15.6	70	15	21.4
3	712	115	16.2	715	129	18.0
4	745	111	14.9	391	83	21.2
5	27	2	7.4	26	2	7.7
6	500	48	9.6	499	56	11.2
	2106	301		1748	303	

Chalmers *et al*. pointed out that the *arithmetic average* of the mortality percents over the six trials was 15.4 for anticoagulated patients and 19.6 for controls.

a. Using the above table, compute the *crude* mortality percent for each group separately and compare your results with the values obtained by Chalmers.

b. How do your answers to part (a) differ from those obtained by by Chalmers *et al*. in terms of the *weights* used for averaging (i.e., what are the weights for each averaging process)?

c. Assuming that your objective is to provide a summary estimate of the *difference in mortality percents*, describe the weighting scheme you would recommend. Explain the reason for your choice. (No calculations are necessary).

d. Assuming that pooling the results of the six trials is appropriate from a study design standpoint, describe (without calculations) how you would carry out a test of hypothesis to determine whether there is a significant "overall" effect of anticoagulant therapy which pools the results of all six trials.

e. Describe (without calculations) how you would compute a 95% interval estimate of the overall difference in mortality percents.

f. How would you criticize, if at all, the use of any kind of pooling procedure over the different trials?

Q.4 A. Describe the processing of the DATA step listed below. Your answer should indicate a description of both the compilation and the execution phase, and a detailed description of the SAS data set created. Your answer should include the following terms/concepts:

input buffer	length
compilation	type
execution	data matrix
program data vector	missing values
variable name	EBCDIC/floating point representation
history information	

```
DATA WORK.ONE;
     LENGTH NAME $ 10 QUIZ1-QUIZ3 4;
     INPUT NAME $ 1-10
           QUIZ1 12-14
           QUIZ2 16-18
           QUIZ3 20-22
           ;
AVERAGE=MEAN(QUIZ1, QUIZ2, QUIZ3);
OUTPUT;
RETURN;
CARDS;
```

B. Two SAS datasets, A.B and A.C are listed on the following pages. Each part of this problem specifies a DATA step which inputs one or both of these two datasets and creates an output SAS dataset. The purpose of these problems is to evaluate your understanding of how statements such as SET, MERGE, UPDATE, work, with and without BY statements. We purposefully made the datasets large

to give you an opportunity to demonstrate YOUR knowledge -- we have confidence that SAS knows how these DATA steps work!

Specify all answers to the number of digits shown in the printout(s). Note that both A.B and A.C are sorted BY ID DATE.

1.1 Step:

```
DATA ONE;
    SET A.B A.C;
    BY ID;
```

Your question: What is the value of WGT in the 17th observation of WORK.ONE?

1.2 DATA Step:

```
DATA TWO;
    SET A.B A.C;
```

Your question: What is the value of WGT in the 17th observation of WORK.TWO?

1.3 DATA Step:

```
DATA THREE
    MERGE A.B A.C;
```

Your question: What is the value of WGT in the 17th observation of WORK.THREE?

1.4 DATA Step:

```
DATA FOUR;
    UPDATE A.B A.C;
    BY ID;
```

Your question: What is the value of WGT in the 17th observation of WORK.FOUR?

1.5 DATA Step:

```
DATA FIVE
    MERGE A.B A.C;
    BY ID DATE;
```

Your question: What is the value of WGT in the 17th observation of WORK.FIVE?

DATASET A.B

OBS	DATE	DATEBORN	ID
1	20SEP78	24APR75	895
2	27FEB78	19MAR75	898
3	08MAR78	23FEB75	899
4	27FEB78	10DEC74	900
5	03MAY78	18MAR74	902
6	05APR78	05MAR74	903
7	05APR78	04MAR74	904
8	15SEP76	22NOV73	906
9	05APR78	25OCT73	907
10	29MAR78	04DEC73	908
11	08DEC76	01DEC73	909
12	20OCT76	10SEP73	911
13	15SEP76	04SEP73	912
14	16SEP76	14JUL73	913
15	16SEP76	14JUL73	914
16	14SEP77	18JUL73	916
17	15SEP76	17JUN73	917
18	16OCT75	18JAN73	919
19	16JAN76	05DEC72	920
20	08DEC76	27NOV72	921

DATASET A.C

OBS	ID	DATEBORN	DATE	WGT	HT	FVC	RACE	SEX	PEF
1	898	19MAR75	05APR78	14	93	0.50	B	M	1.29
2	899	23FEB75	16OCT79	15	95	0.63	B	F	2.85
3	903	05MAR74	20SEP78	17	104	0.81	B	M	3.17
4	904	04MAR74	06SEP78	16	100	0.76	B	M	2.30
5	906	22NOV73	05JAN77	15	99	0.52	B	F	2.40
6	906	22NOV73	07APR77	16	102	0.66	B	F	1.76
7	906	22NOV73	07JUL77	16	104	0.65	B	F	2.05
8	906	22NOV73	11JAN78	18	107	0.74	B	F	1.81
9	906	22NOV73	09AUG78	19	110	0.97	B	F	3.04
10	906	22NOV73	02NOV78	20	110	0.94	B	F	2.94
11	907	25OCT73	11AUG78	17	104	0.96	B	F	2.73
12	907	25OCT73	02NOV78	17	104	1.02	B	F	2.44
13	908	04DEC73	28JUN78	17	104	0.73	B	F	2.65
14	908	04DEC73	06OCT78	18	104	0.86	B	F	2.48
15	909	01DEC73	18MAR77	13	95	0.48	B	M	1.83
16	909	01DEC73	07JUL77	13	99	0.55	B	M	2.11
17	909	01DEC73	12OCT77	14	100	0.60	B	M	1.79
18	909	01DEC73	11JAN78	14	100	0.58	B	M	1.89
19	909	01DEC73	17APR78	15	102	0.58	B	M	2.56
20	911	10SEP73	05JAN77	15	99	0.70	B	F	2.42
21	911	10SEP73	06APR77	16	101	0.73	B	F	2.67
22	911	10SEP73	07JUL77	16	103	0.77	B	F	1.34
23	911	10SEP73	12OCT77	17	105	0.81	B	F	2.21
24	911	10SEP73	11JAN78	17	107	0.80	B	F	2.45
25	911	10SEP73	17APR78	19	108	0.98	B	F	2.88
26	911	10SEP73	24JUL78	19	109	1.03	B	F	2.92

(The dataset A.C is continued on the next page.)

DATASET A.C continued

OBS	ID	DATEBORN	DATE	WGT	HT	FVC	RACE	SEX	PEF
27	912	04SEP73	08DEC76	13	92	0.38	B	M	1.11
28	912	04SEP73	07JUL77	13	96	0.46	B	M	1.44
29	912	04SEP73	26JAN78	15	99	0.54	B	M	1.35
30	912	04SEP73	01MAY78	15	101	0.59	B	M	1.52
31	912	04SEP73	06NOV78	17	103	0.66	B	M	1.94
32	913	14JUL73	11MAY77	16	100	0.65	B	M	1.85
33	913	14JUL73	06FEB78	18	107	0.80	B	M	2.00
34	914	14JUL73	09NOV76	16	97	0.75	B	M	1.33
35	914	14JUL73	05JAN77	16	100	0.80	B	M	1.50
36	914	14JUL73	11MAY77	16	102	0.76	B	M	1.75
37	916	18JUL73	12OCT77	19	107	0.88	B	M	2.27
38	916	18JUL73	13JAN78	19	107	0.81	B	M	1.94
39	916	18JUL73	26JAN78	19	108	0.84	B	M	1.83
40	916	18JUL73	08MAY78	20	109	0.93	B	M	3.02
41	917	17JUN73	14MAR77	15	96	0.79	B	F	2.49
42	917	17JUN73	13JUL77	16	100	0.83	B	F	2.48
43	917	17JUN73	12OCT77	17	102	0.84	B	F	2.53
44	917	17JUN73	11JAN78	17	103	0.85	B	F	2.77
45	917	17JUN73	17APR78	18	104	0.92	B	F	3.22
46	917	17JUN73	24MAY78	18	104	0.91	B	F	3.80
47	917	17JUN73	13SEP78	18	105	1.11	B	F	3.19
48	917	17JUN73	01NOV78	18	105	1.18	B	F	3.38
49	919	18JAN73	18NOV75	15	98	0.56	B	F	1.50
50	919	18JAN73	30MAR76	15	99	0.67	B	F	1.37
51	919	18JAN73	08JUL76	17	104	0.70	B	F	1.88
52	919	18JAN73	14SEP76	17	104	0.78	B	F	2.06
53	919	18JAN73	31JAN77	19	108	0.79	B	F	2.02
54	919	18JAN73	28APR77	19	110	0.87	B	F	2.07
55	919	18JAN73	13JUL77	19	112	0.98	B	F	2.15
56	919	18JAN73	01MAR78	21	114	1.03	B	F	1.93
57	919	18JAN73	27SEP78	22	119	1.17	B	F	2.69
58	919	18JAN73	01NOV78	22	119	1.15	B	F	2.60
59	920	05DEC72	30MAR76	14	94	0.38	B	M	1.46
60	920	05DEC72	06MAY76	14	94	0.45	B	M	1.85
61	920	05DEC72	07JUL76	14	96	0.46	B	M	1.79
62	920	05DEC72	01DEC76	14	99	0.46	B	M	1.46
63	920	05DEC72	31JAN77	15	99	0.52	B	M	1.60
64	920	05DEC72	28APR77	15	101	0.76	B	M	2.24
65	920	05DEC72	12OCT77	16	103	0.80	B	M	1.91
66	920	05DEC72	13JAN78	16	105	0.80	B	M	2.10
67	921	27NOV72	23MAR77	18	108	0.73	B	M	1.72
68	921	27NOV72	13JUL77	19	110	0.80	B	M	2.31
69	921	27NOV72	14SEP77	20	112	0.99	B	M	1.56
70	921	27NOV72	12OCT77	20	112	1.00	B	M	1.44
71	921	27NOV72	26JAN78	23	114	0.88	B	M	1.92
72	921	27NOV72	19APR78	24	114	1.04	B	M	2.48
73	921	27NOV72	24JUL78	24	119	1.17	B	M	2.63

DEPARTMENT OF BIOSTATISTICS

Special MS Examination, 1983
April 9, 1983: 1 - 4 PM

Part I

Answer any three of the following questions. This is a closed book examination.

Q. 1. A city transportation department is conducting a survey to determine the gasoline usage of its residents. Stratified random sampling is used and the four city wards are treated as the strata. The amount of gasoline purchased in the last week is recorded for each household sampled. The strata sizes and the summary information obtained from the sample are:

	STRATA			
	I	II	III	IV
Stratum size	3750	3272	1387	2475
Sample size	50	45	30	30
Sample mean (in gallons)	12.6	14.5	18.6	13.8
Sample variance	2.8	2.9	4.8	3.2

a) Estimate the mean weekly gasoline usage per household for the city population and construct a 95% error bound for your estimate.

b) Estimate the width of a 95% error bound for the estimator in simple random sampling (ignoring stratification when $n = 155$).

c) Do you agree with the surveyor that stratification has led to some reduction in the sampling error of the estimation (over simple random sampling)?

Q.2. For the exponential distribution having probability density function

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x > 0,$$

show that the maximum likelihood estimator of μ is the mean $\overline{X}$ of a random sample of size n from the population $f(x)$. Is this an unbiased estimator of μ? When n is reasonably large, the distribution of $\overline{X}$ can be satisfactorily approximated by a normal distribution. Using this approximation, derive a $100(1-\alpha)\%$ confidence interval for μ. Apply the result to obtain a 95% confidence interval when a sample of 49 yields $\overline{X} = 11.52$. Also, estimate the probability that X is greater than 15.

Q.3. In an experiment designed to determine the relationship between the dose of a compost fertilizer x and the yield of a crop y, the following summary statistics are recorded:

$$n = 15 \qquad \bar{x} = 10.8, \qquad \bar{y} = 122.7$$
$$S_x^2 = 70.6, \qquad S_y^2 = 98.5, \qquad S_{xy} = 68.3$$

Assume a linear relationship.

a) Find the equation of the least squares regression line.

b) Compute the error sum of squares and estimate σ^2.

c) Do the data contradict the experimenter's conjecture that over the range of x values covered in the study, the average increase in yield per unit increase in the compost dose is at least 1.5?

d) Construct a 95% confidence interval for the expected yield corresponding to $x = 12$.

Q.4. Consider the 2×2 contingency table with fixed row totals

	B_1	B_2	
A_1	n_{11}	n_{12}	n_{10}
A_2	n_{21}	n_{22}	n_{20}
	n_{01}	n_{02}	n

Let p_1 and p_2 denote $P(B_1|A_1)$ and $P(B_1|A_2)$, respectively. To test $H_0: p_1 = p_2$, we can employ the normal test with the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(1/n_{10}) + (1/n_{20})}}$$

where $\hat{p}_1 = n_{11}/n_{10}$, $\hat{p}_2 = n_{21}/n_{20}$, and $\hat{p} = n_{01}/n$. Prove that Z^2 is exactly the same as the χ^2 statistic.

b) Prove that the formula for the χ^2 statistic for a 2×2 contingency table can also be written

$$\chi^2 = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{10}n_{20}n_{01}n_{02}} .$$

[Note that χ^2 is the Pearsonian goodness of fit test-statistic.]

SPECIAL MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

April 10, 1983

(1 PM - 4 PM)

INSTRUCTIONS:

a) This is an open-book "in class" examination.

b) Answer from Part I any two of the 3 questions which follow. Also answer Q.4 from Part II. (Thus 3 answers in all)

c) Put the answers to different questions on separate sets of papers.

d) Put your *code letter*, not your name on each page.

e) Return the examination with a signed statement of the honor pledge on a page separate from your answer.

PART I (of PART II)

Q.1 Assume that n=100 individuals participate in a study. A response variable Y is measured, along with a continuous factor X_1 and the presence or absence of a factor X_2. The individuals come from four groups (A,B,C,D), coded as X_3.

a) Suppose that a regression model is fit, predicting Y from the three main effects, all two-way interactions, and the three-way interaction. Fill in the summary ANOVA table and the degrees of freedom in the detailed table. The second table gives the extra sums of squares as each factor is added to the model (the SAS Type I sums of squares).

Source	df	SS	MS	F
Model	____	1500	____	____
Error	____	____	____	
Corrected total	____	2340		

Source	df	Type I SS
X_1	____	800
X_2	____	200
X_3	____	330
X_1X_2	____	50
X_1X_3	____	60
X_2X_3	____	30
$X_1X_2X_3$	____	30

b) Show that the three-way interaction term can be deleted from the model.

c) Test whether all the two-way interaction terms could simultaneously be deleted from the model not containing the three-way interaction.

d) Assume that X_2 is coded as 0 (absent) or 1 (present), and that a model is fit using only the data from group A. This model contains main effects for X_1 and X_2, and the X_1X_2 interaction term, with parameter estimates for X_1,X_2 and X_1X_2 of -2, +10, and +4, respectively. The estimate of the overall mean is 50.

i. What is the predicted value of Y for a subject in Group A with X_1=10 if X_2 is absent? If X_1=10 and X_2 is present?

ii. What is the predicted change in Y if X_1 changes from 10 to 5 in an individual in Group A with X_2 present?

e) Suppose that the mean values for Y for X_2 present and absent, in each group, are as shown in the following table. Assess informally whether there is an interaction between group and the presence or absence of X_2. Do not do an hypothesis test.

Mean Values of Y

Group	X_2 Absent	X_2 Present
A	30	50
B	25	15
C	35	45
D	40	40

Q.2

a) Some or all of the JCL statements below contain one or more errors. Circle each error, the whole error, and nothing but the *error*, and briefly explain what is wrong. Consider each statement separately. Assume that blanks within the lines occur *only* where the symbol "_" appears. None of the errors involves "1" versus "l" or "0" versus "o".

```
                              COLUMN RULER
0000000001111111111222222222233333333334444444444555555555566666666667 77
12345678901234567890123456789012345678901234567890123456789012345678901 2
```

```
(1).  //_JOB_UNC.B.E1234,HOSKING

(2).  //_EXEC_PGM=COPY,WASTE=YES

(3).  //PHRED_DD_UNIT=DISK,DISP=(NEW,DELETE),VOL=REF=UNCCC.OFFLINE,
      //_________SPACE=(TRK,(10,20)),DCB=(RECFM=FB,BLKSIZE=600,LRECL=250),
      //_________DSN=UNC.B.E1234.JONES.X

(4).  //COPYDISKS_JOB_UNC.B.E1234,BROWN,T=(,45),M=0

(5).  //EXEC_PGM=COPY,PARM=LIST

(6).  //OUTPUT_DD_DSN=UNC.B.E9999.JONES.MYSTUFF,UNIT=TAPE,DISP=(NEW,CATLG),
      //_RING=IN,LABEL=(3,SL)

(7).  //JOE_DD_UNIT=DISK,DSN=UNC.B.E1234.DATA.407,DISP=OLD,DCB=(RECFM=FB,
      //____LRECL=80,BLKSIZE=6000)

(8).  //SYSOUT_DD_DSN=UNC.B.E1234.HELMS.STUFF,
      //_______VOL=REF=UNCCC.ONLINE,
      //_______DCB=(RECFM=VB,BLKSIZE=6000,LRECL=5996),
      //_______SPACE=(TRK,(10,5,RLSE)),UNIT=DISK,
      //_______DISP=(NEW,CATLG,DELETE)

(9).  //INPUT_DD_DSN=UNC.B.E5001.SMITH.INDATA.ONTAPE,LABEL=(3,SL)
      //_________DISP=OLD

(10).  //JOB#1_JOB_UNC.B.E1234_SMITH,REGION=999K,MSGLEVEL=(1,0),TIME=5
```

b) Describe the processing of the DATA step listed below. Your answer should indicate a description of both the compilation and the execution phase, and a detailed description of the SAS data set created. Your answer should include the following terms/concepts:

input buffer	length
compilation	type
execution	data matrix
program data vector	missing values
variable name	EBCDIC/floating point representation
history information	

```
DATA WORK.ONE;
     LENGTH NAME $ 10 QUIZ1-QUIZ3 4;
     INPUT NAME $ 1-10
           QUIZ1 12-14
           QUIZ2 16-18
           QUIZ3 20-22
           ;
AVERAGE=MEAN(QUIZ1, QUIZ2, QUIZ3);
OUTPUT;
RETURN;
CARDS;
```

Q.3 For the period 1 July 1974 through 30 June 1978, North Carolina experienced a sudden infant death (SID) rate of two per thousand live births.

a) For a county having 3000 live births and 12 SIDS in the same calendar period, calculate the P-value for the possibility that the SID rate in this county is greater than that of North Carolina. By analogy to a binomial situation, use the normal distribution to approximate a presumed underlying Poisson model.

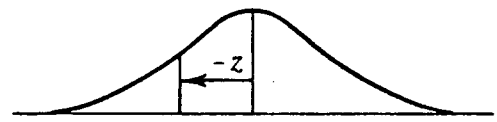
b) Let λ denote the SID rate for a county with 3000 live births during the above specified calendar period and consider the null hypothesis $Ho:\lambda=0.002$, i.e., λ is two SIDS per thousand live births. Determine an upper tail critical region with a level of significance $\alpha=0.05$. Again use a normal approximation.

c) Following part (b), calculate the power corresponding to the alternative hypothesis $Ha:\lambda=0.006$.

d) Comment briefly on the appropriateness of the normal approximation to the Poisson distribution for the above calculations. Sketch how you would proceed in parts (a), (b), and (c) if the normal approximation were not appropriate.

APPENDIX 609

Table I. *Normal Probability Distribution Function (Probabilities That Given Standard Normal Variables Will Not Be Exceeded—Lower Tail)* $N_z(-z)$. Also $N_z(z) = 1 - N_z(-z)$.



$-z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207

EDITORIAL NOTE. The table above has been abridged from one extended to Z = 3.99 which was attached to the original examination.

PART II (of PART II)

Q.4 A clinical trial is conducted to assess the efficacy of a new drug to alleviate symptoms of depression. Patients are randomized to Drug (D) or Placebo (P) in equal proportions at three psychiatric clinics. A total of 120 patients are entered with 20 patients allocated to each treatment group at each clinic. Patients are tested for depression at baseline and at weeks 1, 2, 3, and 4. About 30% dropout occurs by week 4 (this attrition is to be expected in depression trials) and so, the statistical analyst decides to concentrate on "final rated value" as the main dependent variable. The scale of measurement of major interest is Total Score of Hamilton Depression Scale. This scale is from 0-62 with

0-13 ~ little or no depression (essentially "cured")

14-19 ~ minor depression (not sick enough to enter trial but not well enough to be "cured")

20-29 ~ moderate depression

$\geq$ 30 ~ severe depression.

Some controversy ensues about the univariate statistical analysis of the final rated values of the Hamilton Depression Scale. One aspect of the controversy is that there is significant treatment x clinic interaction. Clinic 1 shows a preference for placebo (non-significant); clinic 2 shows a preference for drug (non-significant); clinic 3 shows a significant preference for drug. The other aspect of the controversy is whether to adjust for the pre HAMD score via covariance analysis. In clinic 1, placebo patients are a little more severe at baseline (not significant, $p = .19$); in clinic 2, there are no treatment group differences at baseline ($p = .56$); in clinic 3, drug patients are more severe at baseline (near significance, $p = .08$). Six different methods of analysis to test treatment effects are proposed by different statisticians. These are as follows:

1. Two-way ANOVA

   Two-way ANOVA of final HAMD scores employing treatment and clinic as main effects and treatment by clinic interaction.

2. Two-way ANOVA of differences

   Two way ANOVA of "final HAMD score minus pre HAMD score" employing treatment and clinic as main effects and treatment x clinic interaction.

3. Two-way ANCOVA

   ANCOVA of final HAMD scores employing pre HAMD score as a covariate. The main effects treatment and clinic would be included in the model and so would treatment x clinic interaction.

4. Separate ANOVA's

   A t-test on final HAMD scores for each clinic separately.

5. Separate ANCOVA's--different slopes

   ANCOVA on final HAMD scores for each clinic separately. The covariate is pre HAMD score and the main effect is treatment (i.e., covariance adjusted t-test).

6. Separate ANCOVA's--common slope

   ANCOVA on final HAMD scores for each clinic separately. The (linear) covariate is pre HAMD score but with a slope that is derived from a pooled analysis of all three clinics. The main effect is treatment.

Question

(18 points) i. Comment on each of the proposed analyses outlining the advantages and disadvantages of each.

(7 points) ii. Which of the six analysis strategies would you choose? Explain why.

NOTE: For the purposes of this question, assume that the usual parametric assumptions of normally and independently distributed errors with zero mean and constant variance are reasonably appropriate. In other words, do not deal with the issues regarding whether a non-parametric or parametric analysis is appropriate.

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(January 21, 1984)
9:30 AM to 12:30 PM

1. Suppose that the main purpose of a statewide household survey is to estimate the proportion (P) of North Carolina households without any form of health insurance. A self-weighting household sample is selected by using simple random sampling in each of three stages, with clusters of unequal size in the first two stages. The estimator of P is the simple proportion of uninsured households in the sample. The estimate turns out to be 0.24, with a standard error of 0.02 and a design effect of 1.2.

   (a) State whether or not the estimator of P is unbiased. Briefly explain your answer.

   (b) How many sample households produced the estimate (i.e., 0.24)?

   (c) How many sample households, chosen by simple random sampling (i.e., no cluster sampling), would have been needed to produce the same amount of statistical precision (i.e., standard error of 0.02)?

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu,\sigma^2)$. Let $S^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2$. Using the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad ,$$

solve the following problems:

   (a) Find $E(S^2)$ and $V(S^2)$.

   (b) Find random variables A and B such that $pr(A<\sigma^2<B)=(1-\alpha)$, $0<\alpha<1$.

   (c) Let $t>0$ and $0<p<1$. Use Tchebysheff's Theorem to find the smallest sample size n such that

$$pr\{|S^2-\sigma^2|<t\sigma^2\}\geq p.$$

   [<u>HINT</u>: Your lower bound for n will be a function of t and p.]

3. Let Y be normally distributed with mean μ and variance σ^2. A random sample of size n is drawn of values of Y. The sample mean is used to estimate μ and to test $H_o:\mu=0$. For parts (a) and (b), assume that σ^2 is known.

Suppose that detection of a fixed alternative $\mu=m$ 0 is of interest, and that the hypothesis test is done with $\alpha=.05$.

(a) Determine the chance of rejecting H_o (the power of the test) if the sample size is chosen so that the half-width of a 95% confidence interval for μ is m.

(b) Suppose that sample size is determined to obtain a given power for the fixed alternative $\mu=m$. Show that, with this sample size, the half-width of a 95% confidence interval for μ is a fixed percent of m, the percent depending on the power but not on m and not on σ^2.

(c) Show how that analysis is modified if the variance is unknown. Give a very brief proof that your result in (a) is still valid.

4. Suppose that the length of life of electric tube, T, has an exponential probability density function

$$f_T(t)=\lambda e^{-\lambda t},\ t>0,\ \lambda>0.$$

In a sample of n tubes observed for a period τ, d tubes failed, while the lifetimes of the remaining tubes were greater than τ.

(a) Find the maximum likelihood estimator of λ ($\hat{\lambda}_n$, say).

(b) Find the approximate expected value of $\hat{\lambda}_n$.

(c) Show that the approximate variance (its lower bound) of $\hat{\lambda}_n$ is $\mathrm{Var}(\hat{\lambda}_n) = \frac{1}{\tau^2 n}(e^{\lambda\tau}-1)$.

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(January 22, 1984: 2-5 PM)

1. Assume that the dataset UNC.B.E555V.BIOSTAT.LIPID.LAB contains data records keyed from the attached Lipid Laboratory Data Form. Assuming this is a catalogued dataset stored at TUCC, write the program needed to create a permanent SAS dataset from this file, perform some basic edits, and print some descriptive statistics. This program should consist of the appropriate JCL and three SAS steps which perform the following tasks:

   a) Read the data records and create a SAS dataset named LABONE. This SAS dataset should be stored in a SAS data library created by your job and named

      UNC.B.E555V.BIOSTAT.LIPID.SASDATA

      Read each of the fields except those identified as "always blank." Use the SAS variable names shown in the attached table. Provide appropriate formats and labels for these variables. This portion of the program should include JCL statements to read the data file and create the SAS data library. Assume the following JCL statements are given:

```
//EXAM JOB UNC.B.E123X,STUDENT
//*PW=LUCK
// EXEC SAS
```

   b) Write a SAS DATA step to read LABONE and create a temporary SAS dataset named LABCLEAN containing only those data values that pass the following edit tests:

variable	valid values
FORMNO	SLDA1
CHYLO	1, 2, 9
TRIG	0-2000
TRIGBLK	0-99
TRIGADJ	TRIG-TRIGBLK (i.e., the value of TRIG minus the value of TRIGBLK)

      This step should print an error message on the SAS log, with appropriate information, for each data value which fails an edit and then set the value to missing.

   c) Write a SAS PROC step to print descriptive statistics (including the number of missing values, mean, minimum value, maximum value, and variance) for the variables TRIG and TRIGADJ. Provide appropriate titling information.

Field Specification for the Lipid Lab Form

Variable	Columns	Format[1]
FORMNO	1-5	AAAAN
	6-16	always blank
VISDATE	17-22	MMDDYY
LASTNAME	23-34	AAA.....
LASTINIT	23	A
INITIALS	35-36	AA
CHYLO	37	N
	38-54	always blank
TRIG	55-58	NNNN
TRIGBLK	59-60	NN
TRIGADJ	61-64	NNNN
TRIGDATE	65-70	MMDDYY
	71-80	always blank

[1] A indicates byte can contain letters, numerals, or special characters
N indicates byte can contain only numerals
MMDDYY indicates dates where MM=month
DD=day
YY=year

Version 73165

LRC PREVENTION TRIAL

LIPID LABORATORY DATA FORM

S L D A 1

(1-5)

1. Date of Visit: Month Day Year (17-22)

2. Last Name: (23-34)

Initials: 1st 2nd (35-36)

3. Standing plasma test:

a. Chylomicron layer:

1 Present . . . . . .

2 Absent . . . . . .

9 Not done . . . . . .

3.

a.

1

2

9 (37)

5. a. Triglycerides *(Record in mg%)*: mg% (55-58)

b. Triglyceride blank *(To be done only if triglyceride value is greater than 300 mg%)*: mg% (59-60)

c. Triglyceride less blank: *Net triglycerides* mg% (61-64)

d. Date of triglyceride determination: Month Day Year (65-70)

2 Femoral antiversion (toeing in) is customarily detected and measured by X-ray technique. It is desired to minimize the use of X-rays because it exposes the gonadal area to possible hazard. Thus variation in X-ray technique to minimize X-ray exposure is a common object of study. The following data resulted in one such study in which various degrees of antiversion were produced in a model and then read on X-ray film by three technicians. Each technician read each film twice in random order.

		Technician		
		1	2	3
Angle of antiversion in model	0	5,2	3,3	-8,-2
	15	18,17	17,16	18,15
	30	34,34	32,35	34,34
	45	46,46	44,46	43,45
	60	66,61	60,59	48,56

1) Investigate the bias in the new method.

2) Do the three technicians differ more than expected by chance?

3) Do the technicians' biases vary with the angle?

4) Write a 1/2 page report summarizing your findings.

3 Data concerning the amount of heat evolved in calories per gram of cement (Y) as a function of the amount of each of four ingredients (X1, X2, X3, X4) in the mix are presented in Table 1. Be brief and to the point in your response to each part.

(a) An all possible regression functions analysis is reported in Table 2. Interpret the results of this analysis.

(b) A stepwise regression analysis is reported in Table 3. Interpret the results of this analysis.

(c) How many variables would you include in a regression model? Which are they?

(d) What have you learned from this problem about setting the entry and exit criteria when performing stepwise regression on a data set too large for an all possible regression functions analysis?

(e) Describe what other analyses you would perform to assess whether the assumptions underlying regression analysis are met, for the regression model chosen in (c).

(f) How would you use the regression model selected in (c) to predict the amount of heat that is likely to evolve from yet another mix of cement?

Table 1: Data

Observation, i	(CAL) y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Table 2: Summary of all possible regressions

Number of Regressors in Model	p	Regressors in Model	$SS_E(p)$	R_p^2	$\bar{R}_p^2$	$MS_E(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Table 3

FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP 1 VARIABLE X4 ENTERED R SQUARE = 0.67454196
C(P) = 136.73063349

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	1	1831.89616002	1831.89616002	22.80	0.0006
ERROR	11	883.86691690	80.35153790		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	117.56793118				
X4	-0.73616181	0.15459600	1831.89616002	22.80	0.0006

STEP 2 VARIABLE X1 ENTERED R SQUARE = 0.97247105
C(P) = 5.49585082

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	2	2641.00096477	1320.50048238	176.63	0.0001
ERROR	10	74.76211216	7.47621122		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	103.09738164				
X1	1.43995828	0.13641664	809.10480474	108.22	0.0001
X4	-0.61395363	0.04864455	1190.92483664	159.30	0.0001

STEP 3 VARIABLE X2 ENTERED R SQUARE = 0.98233545
C(P) = 3.01823347

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	3	2667.79034752	889.26344917	166.83	0.0001
ERROR	9	47.97272940	5.33030327		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	71.64830697				
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	-0.23654022	0.17328779	9.93175378	1.86	0.2054

STEP 4 ALL VARIABLES ENTERED R SQUARE = 0.98237502
C(P) = 5.00000000

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	4	2667.89943757	666.97485939	111.48	0.0001
ERROR	8	47.86363935	5.98295492		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	62.40536930				
X1	1.55110265	0.74470967	25.95091138	4.34	0.0708
X2	0.51016758	0.72378600	2.97247824	0.50	0.5009
X3	0.10190940	0.75470905	0.10909005	0.02	0.8959
X4	-0.14406103	0.70905206	0.24697472	0.04	0.8441

4 In a retrospective study of the possible effect of blood group on the incidence of peptic ulcers, Woolf (1955) obtained data from three cities. The table below gives for each city data for blood groups O and A only. In each city, blood group is recorded for peptic-ulcer subjects and for a control series of individuals not having peptic ulcer.

Blood groups for peptic ulcer and control subjects

	Peptic ulcer		Control	
	Group O	Group A	Group O	Group A
London	911	579	4578	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

Source: Woolf (1955). On estimating the relation between blood group and disease. Annals of Human Genetics, 19: 251-3.

(a) Investigate the association between blood group and peptic ulcer status at each locality. Use a confidence interval method and interpret its implications to the hypothesis of no relationship.

(b) Assess the homogeneity of any blood group by peptic ulcer status association across localities. Use an appropriate significance test. Also, if homogeneity is supported, provide a confidence interval for the blood group by peptic ulcer status association for the combined data from all three localities.

(c) Discuss briefly the interpretation of the results from (a) and (b).

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(April 15, 1984)

INSTRUCTIONS:

a) This is an *open book* examination.

b) M.P.H. students are to answer *any two* questions during the two-hour period (1:30 pm - 3:30 pm). M.S. students are to answer *three* questions of which not more than 2 should be from Group A - time period 1:30 pm - 4:30 pm.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, *not* your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

f) You are required to answer *only what is asked* in the questions and *not all you know* about the topics.

Group A

1. A randomized clinical trial for hypertension is to be designed to compare four treatment modalities in a parallel group design. The physician in charge of the trial wants to know the correct sample size to use. Review of past trials shows that the standard deviation for diastolic blood pressure phase V of entering patients is around 6 mm Hg. The physician feels that differences between treatment groups of 3 mm Hg must be detectable by the experiment.

Perform some power calculations using the attached charts to help identify appropriate sample sizes for the following situations:

(I) If any two treatments differ by 3 mm Hg, then the 4 treatment group comparison with 3 d.f. for the numerator of the F-test should be significant.

(II) If any two treatments differ by 3 mm Hg, then the pairwise difference based on the t-test should be significant.

(III) If any two treatments differ by 3 mm, then a Bonferroni adjusted pairwise difference (based on the fact that six pairwise differences are being inspected) should be significant.

EDITORIAL NOTE. The "attached charts" were those given on pages 115 and 116 in E.S. Pearson and H.O. Hartley, "Charts of the power function for analysis of variance tests, derived from the non-central F-distribution", Biometrika 38 (1951) 112-130.

2. A teacher wishes to determine the value of providing a manual and/or certain notes to his classes. He has 48 students, whom he distributes at random among 4 different groups, placing 12 in each. The assignment of teaching aid combinations to groups is also done at random. After the course is over, all students who are still enrolled take the same exam, with results as shown below.

Manual	Notes	N	Data	Mean	S.D.
Yes	Yes	9	60,64,67,68,68,69,71,73,75	68.33	4.53
Yes	No	10	32,41,44,47,48,54,54,64,65,73	52.20	12.42
No	Yes	11	41,44,47,49,51,54,56,59,61,61,73	54.18	9.16
No	No	7	44,51,55,59,59,66,69	57.57	8.56

ANOVA:

Source	DF	SS	MS
Between	3	1456	485.3
Within	33	2831	85.8
Total	36	4287	119.1

(These data are from W.C. Guenther, Analysis of Variance, page 44.)

a) Assuming that differences in average grade may be attributed solely to the different combinations of teaching aids, write out a linear model for this experiment. Estimate all its parameters.

b) Consider two simplifications of the model:

1) Eliminate interaction, making it additive.
2) Combine all groups except Yes/Yes.

Show that the data will support the second simplification but not the first.

c) Present an argument for transforming these data. What would be a good transformation to try? (Do not actually perform any analysis with transformed data.)

d) Criticize the experimental design, and/or the use of standard analysis of variance with it. Can you suggest improvements?

3. Your answers to Question 3 must be submitted on these sheets. NO MATERIAL ON ANY OTHER SHEET WILL BE CONSIDERED!

You may use the back of the sheet if necessary, but enough space has been left under each part to provide the expected answer for that part.

Parts (a)-(e) of this question are worth three points each. In each case, the DATA step shown is executed reading one or both of the SAS data sets shown below (note that both are sorted BY ID). For each step, list the data part of the data set created (include variable names as shown below), and answer the questions presented.

data set ONE

ID	X	Y
1	4	•
1	10	8
3	•	4
4	2	5

data set TWO

ID	X	Z
1	•	-3
2	6	0
3	7	9
3	4	•

Output Dataset

a)
```
DATA A;
  SET TWO;
  Y=SUM(X,Z);
  IF (X LT Z) THEN OUTPUT;
  RETURN;
```

How many times is the DATA step executed? _____.

data set ONE

ID	X	Y
1	4	.
1	10	8
3	.	4
4	2	5

data set TWO

ID	X	Z
1	.	-3
2	6	0
3	7	9
3	4	.

b)
```
DATA B;
  SET TWO ONE(RENAME=(Y=Z));
  IF (X GE 4);
  Q=X+Z;
  DROP X;
  OUTPUT;
  RETURN;
```

Output Dataset

What variables are in the PDV?

c)
```
DATA D;
  MERGE TWO ONE;
  BY ID;
  IF LAST.ID THEN OUTPUT;
  RETURN;
```

Output Dataset

How many observations would be in the data set if the OUTPUT statement was replaced by a DELETE statement (following the IF/THEN)?

data set ONE

ID	X	Y
1	4	•
1	10	8
3	•	4
4	2	5

data set TWO

ID	X	Z
1	•	-3
2	6	0
3	7	9
3	4	•

d)
```
DATA D;
  SET TWO;
  DROP X Z;
  VAR='X';
  VAL=X;
  OUTPUT;
  VAR='Y';
  VAL=Y;
  OUTPUT;
  RETURN;
```

Output Dataset

What is the length of VAL? Why?

e)
```
DATA E;
  SET ONE;
  NUM=PUT(ID,WORDS5.);
  AVG=(X+Y)/2;
  MEAN=MEAN(X,Y);
  DROP ID X Y;
  OUTPUT;
  RETURN;
```

Output Dataset

What is the length of NUM? Why?

Parts (f) and (g) are worth five points each.

In parts (f) and (g) of this problem, you are only asked to compile the descriptor part of the data set, not to execute it or create the data part.

For each of the DATA steps in parts (f) and (g) below, fill in the information SAS would store in the descriptor part of the data sets being created. Sketch the Program Data Vector that would be created, indicating the type and length of each variable in it. When appropriate, indicate the existence and size of input and/or output buffers.

f)
```
DATA F;
  INPUT  NAME $ FNAME $ SATM SATV;
  SATTOT=SATM+SATV;
  IF(SATTOT GT 1000) THEN STATUS='ADMIT';
  ELSE STATUS='REJECT';
  FILE PRINT;
  PUT NAME $ 1-10 STATUS $ 12-20;
  OUTPUT;
  RETURN;
  CARDS;
```

PDV

Name:

Type:

Length:

WORK.D

Name:

Type:

Length:

Informat:

Format:

Label:

g)
```
DATA G;
  INFILE IN LRECL=50;
  INPUT #1 @1  NAME     $20.
           @21 BDATE    MMDDYY8.
           @31 DDATE    MMDDYY8.
        #2 @1  CAUSE1   $6.
           @11 CAUSE2   $6.
         ;
  FORMAT BDATE DDATE DATE.;
  LENGTH SURVTIME 4;
  SURVTIME=DDATE-BDATE;
  OUTPUT;
  RETURN;
```

PDV

Name:

Type:

Length:

WORK.E

Name:

Type:

Length:

Informat:

Format:

Label:

Group B

4. For the period 1 July 1974 through 30 June 1978, North Carolina experienced a sudden infant death (SID) rate of two per thousand live births.

(a) For a county having 3000 live births and 12 SIDs in the same calendar period, calculate the P-value for the possibility that the SID rate in this county is greater than that of North Carolina. Use the normal distribution to approximate a presumed underlying Poisson model.

(b) Let λ denote the SID rate for a county with 3000 live births during the above specified calendar period and consider the null hypothesis $H_o: \lambda = 0.002$, i.e., λ is two SIDs per thousand live briths. Determine an upper tail critical region with a level of significance $\alpha = 0.05$.

(c) Following part (b), calculate the power corresponding to the alternative hypothesis $H_a: \lambda = 0.006$.

(d) Comment briefly on the appropriateness of the normal approximation to the Poisson distribution for the above calculations. Sketch how you would proceed in parts (a), (b), and (c) if the normal approximation were not appropriate.

MS WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(January 26, 1985; 10 AM to 1 PM)

Question 1

A student health service has a record of the total number of eligible students (N) and of the total number of visits (Y) made by students during a year. Some students made no visits. The service wishes to estimate the mean number of visits (Y/N_1) for the N_1 students who made at least one visit, but does not know the value of N_1. A simple random sample of n eligible students is taken in which n_1 students made at least one visit and their total number of visits was y. Ignoring the finite population correction:

a. Show that y/n_1 is an unbiased estimator of Y/N_1.

b. Show that the variance of y/n_1, conditioned on n_1, is S_1^2/n_1, where S_1^2 is the variance of the number of visits among students making at least one visit.

c. An alternative estimator of Y/N_1 would be Yn/Nn_1. Show that the alternative estimator is biased and that the ratio of the bias to the true value (Y/N_1) is approximately $(N - N_1)/nN_1$.

Hint: If $\hat{P}$ is a binomial estimator of $P = 1 - Q$ based on n trials, then

$$E\frac{1}{\hat{P}} = \frac{1}{P} + \frac{Q}{nP^2}$$

Question 2

Consider the density function

$$f_X(x;\ \theta) = \frac{x}{\theta}\, e^{-x^2/2\theta}, \quad x > 0,\ \theta > 0.$$

a) Find E(X) and V(X).

b) Find a sufficient statistic for θ based on a random sample $X_1, X_2, \ldots, X_n$ of size n from $f_X(x;\theta)$.

c) Find an unbiased estimator of θ which is a function of a sufficient statistic for θ. Also, find the variance of this unbiased estimator of θ.

d) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $f_X(x;\theta)$. Find $E[X_{(1)}]$ and $V[X_{(1)}]$, the mean and variance of the smallest observation in this sample.

e) Let $X_1, X_2, \ldots, X_n$ be a random sample from $f_X(x;\theta)$, where n is large. Find an approximate 95% confidence interval for θ which is a function <u>only</u> of $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$.

Question 3

Let X_1 and X_2 be independent exponential random variables with the PDF's

$$F_{X_i}(x_i;\lambda) = \lambda e^{-\lambda x_i},\ \lambda>0,\ x_i>0\ i=1,2.$$

a) Obtain the joint distribution of

$$Y_1 = \frac{X_1}{X_1+X_2} \text{ and } Y_2 = X_1 + X_2$$

b) Obtain the marginal PDF's of Y_1 and Y_2. What known distributions do they represent?

c) Are Y_1 and Y_2 independent?

d) Obtain the rth moment of each of Y_1 and Y_2.

Question 4

Let X be a Poisson random variable with probability mass function

$$P_X(x;\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda},\ \lambda>0,\ x=0,\ 1,2,\ldots.$$

a) Derive the moment of generating function, $M_X(t)$, of X.

b) Obtain the mean and the variance of this distribution.

c) Let $x_1, x_2, \ldots, x_n$ represent an observed sample from this distribution. Construct the likelihood ratio test for the hypothesis $H_0:\lambda=\lambda_0$ against the alternative $H_A:\lambda>\lambda_0$.

d) Obtain a formula for the power of the test developed in (c) for $\lambda=\lambda_1$ $(>\lambda_0)$.

M.S. WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(January 27, 1985: 1:00 to 5:00 PM)

Question 1

An experimenter wished to measure the difference in concentration of tobacco mosaic virus prepared by two methods. For each of 8 leaves, one half was chosen at random for inoculation with virus prepared by the first method, and the other half by the second. The lesions found subsequently were counted as an index of virus concentration in the preparation. The following data resulted.

Leaf Number	Lesions on half inoculated with	
	Preparation 1	Preparation 2
1	9	10
2	17	11
3	31	18
4	18	14
5	7	6
6	8	7
7	20	17
8	10	5

a) Calculate a parametric 95% confidence interval for the true mean difference in number of lesions.

b) What does this tell you about the true mean difference between the two preparations?

c) Select and perform a nonparametric test for the same location in distribution of the number of lesions for the two preparations.

d) Discuss the comparability of the results from parts (b) and (c).

Question 2

Mercury levels (micrograms /gm body weight) were measured in fish at various distance above and below the source of pollution. The data are:

Distance from Source

5.5 km above plant	3.7 km below plant	21 km below plant	133 km below plant
.45	1.64	1.56	.65
.35	1.67	1.55	.59
.32	1.85	1.69	.69
.68	1.57	1.67	.62
.53	1.59	1.60	.70
.34	1.61	1.68	.64
.61	1.53	1.65	.81
.41	1.40	1.59	.58
.51	1.70	1.75	.53
.71	1.48	1.49	.75

a) Provide a descriptive graph of these data.

b) Comment on the reasonableness of the assumptions for a linear regression analyses of these data. How are the data 5.5 km above the plant useful?

c) Provide an estimate of the distance below the plant where the mercury levels should be as they were above the plant.

d) Sketch a line representative of the trend of decreasing mercury levels below the plant.

e) Describe the essential technical difficulties in obtaining a confidence interval for the estimate in part (c) given the analysis in part (d).

Question 3

A random sample of 1000 auto accidents was chosen from records for a state during a year. The accidents were tabulated by type of safety restraint and extent of injury as follows:

Extent of Injury	Type of Restraint: None	Seat Belt Only	Seat Belt & Harness
None	65	75	60
Minor	175	160	115
Major	135	100	65
Death	25	15	10

a) Set up the calculations for a Pearson's chi-square test of independence of extent of injury and type of restraint. What is the corresponding null hypothesis?

b) If that chi-square value is 10.96, approximate its P-value to two decimal places.

c) Describe the attached weighted least squares analyses.

d) What chief result(s) is (are) apparent from parts (b) and (c)?

e) Describe any differences in sampling perspective for the analyses in part (a) and (c).

GENERALIZED CHI-SQUARE ANALYSIS
OF CATEGORICAL DATA

THE DATA ARE READ FROM UNIT 1 .

THE TYPE OF INPUT IS CONTINGENCY TABLE.

THE NUMBER OF SUB-POPULATIONS IS 3 .

THE NUMBER OF RESPONSE PROFILES IS 4 .

CONTINGENCY TABLE:

65.0	175.0	135.0	25.0
75.0	160.0	100.0	15.0
60.0	115.0	65.0	10.0

PROBABILITY VECTOR:

0.16250D 00	0.43750D 00	0.33750D 00	0.62500D-01
0.21429D 00	0.45714D 00	0.28571D 00	0.42857D-01
0.24000D 00	0.46000D 00	0.26000D 00	0.40000D-01

BASIC BLOCK OF LINEAR OPERATOR A1:

0.0	1.00	2.00	4.00

F(P) = A1*P:

0.13625D 01	0.12000D 01	0.11400D 01

COVARIANCE MATRIX:

0.23277D-02	0.0	0.0
0.0	0.24163D-02	0.0
0.0	0.0	0.33616D-02

TOTAL CHI-SQUARE = 1780.0639
DF = 3
P = 0.0

WEIGHTED LEAST SQUARES ANALYSIS

DESIGN MATRIX:

1.00

1.00

1.00

ESTIMATED MODEL PARAMETERS:

0.12455D 01

COVARIANCE MATRIX:

0.87548D-03

STANDARD DEVIATIONS OF THE ESTIMATED MODEL PARAMETERS:

0.29605D-01

CHI-SQUARE DUE TO ERROR = 10.0486 DF = 2 P =0.0066

****

F(P) = A1*P:

0.13625D 01 0.12000D 01 0.11400D 01

F(P) PREDICTED FROM MODEL:

0.12455D 01 0.12455D 01 0.12455D 01

RESIDUAL VECTOR: (F(P) - PREDICTED F(P))

0.11696D 00 -0.45543D-01 -0.10554D 00

STANDARD DEVIATIONS OF THE PREDICTED FUNCTIONS:

0.29605D-01 0.29605D-01 0.29605D-01

```
DESIGN MATRIX:
   1.00     0.0
   1.00     1.00
   1.00     2.00

ESTIMATED MODEL PARAMETERS:
   0.13470D 01   -0.11470D 00

COVARIANCE MATRIX:
   0.19749D-02   -0.12422D-02
  -0.12422D-02    0.14049D-02


STANDARD DEVIATIONS OF THE ESTIMATED MODEL PARAMETERS:
   0.44439D-01    0.37482D-01

CHI-SQUARE DUE TO ERROR =     0.6842          DF =   1          P =0.4081
*****************************************************************************
F(P) = A1*P:
   0.13625D 01    0.12000D 01    0.11400D 01

F(P) PREDICTED FROM MODEL:
   0.13470D 01    0.12323D 01    0.11176D 01

RESIDUAL VECTOR:          ( F(P) - PREDICTED F(P) )
   0.15539D-01   -0.32260D-01    0.22440D-01

STANDARD DEVIATIONS OF THE PREDICTED FUNCTIONS:

   0.44439D-01    0.29922D-01    0.51241D-01
*****************************************************************************


CONTRAST MATRIX:
   0.0      1.00

ESTIMATED MODEL CONTRASTS:
  -0.11470D 00

STANDARD DEVIATIONS OF THE ESTIMATED MODEL CONTRASTS:
   0.37482D-01


CHI-SQUARE =     9.3644          DF =   1          P = 0.0022
```

Question 4

The catalogued OS file UNC.B.E522U.BIOS111.DSPA.RAW contains 2000 data records keyed from the attached Dopler Systolic Blood Pressure Form. These records are each 80 bytes long. Each field is keyed as specified by the column number(s) in parentheses following the boxes or numbers for the data values. Note that not all columns are used since some questions have been eliminated to simplify the form for this exam. Also note that each form generates two records, DSPA1 and DSPA2, as indicated by the record identifiers in columns 1-5. For each part below give the JCL and/or SAS statements necessary to perform the tasks described.

Job 1

Create a catalogued on line dataset named UNC.B.E522U.DSPA.MYCOPY containing a copy of the records in the OS file described above. The output data set should be blocked with 100 logical records (of 80 bytes each) per physical record.

Job 2

Read the data set created in Job 1, creating a SAS data set from it. The SAS data set should have 12 variables, one corresponding to each keyed item on the form. Use the information on the form to choose appropriate variable names, types, lengths and input formats. Permanently store the data set in an online SAS data library created by this job.

Job 3

a) Produce a listing of subjects' last name and initials, sorted by "date form completed", from earliest to latest.

b) Produce a vertical bar chart showing the percentage of forms having 3 bars, one for each of the 3 possible responses to question 10.

c) Produce a scatter plot of right ankle artery pressure versus left (question 7), using the response to question 9 as the plotting symbol. When question 9 is missing, use an asterik (*) as the plotting symbol.

Version 74165

DOPPLER SYSTOLIC PRESSURE FORM

(Prevalence and Prevention Subjects)

O.M.B. 68-S72041 and 68-S73044

D S P A 1 (1-5)

Visit Number (15-16)

I. *Enter the subject's name beginning with the leftmost box; then enter the subject's first and second initials in the boxes labeled "Initials." Enter all letters as capitals.*

1. Subject's Last Name: (17-28)

2. Initials: 1st 2nd (29-30)

3. Date form completed: Month Day Year (31-36)

4. Code number of person completing form: (37-38)

II. SYSTOLIC PRESSURE MEASUREMENTS USING DOPPLER

6. Systolic left arm blood pressure taken in SITTING position prior to treadmill exercising.

Left Arm Blood Pressure: (57-59)

7. Ankle Artery Pressure taken in SITTING position prior to exercising (POSTERIOR TIBIAL MUST BE USED unless unobtainable):

i. Right: (66-68)

ii. Left: (69-71)

NEW CARD D S P A 2 (1-5)

9. *If Dorsalis Pedis used in place of Posterior Tibial in question 7 or 8 above, circle the appropriate response:*

Right artery only 1

Left artery only 2

Both right and left arteries 3 (44)

10. Circle the appropriate response regarding the treadmill exercise test:

Performed and target heart rate reached 1

Performed but target heart rate not reached 2

Not performed 3 (45)

DEPARTMENT OF BIOSTATISTICS

Special MS Examination, 1985

April 13, 1985: 10 AM - 1 PM

INSTRUCTIONS:

a) Answer any three of the following questions. This is a *closed book* examination.

b) Put the answers to different questions on separate sets of papers.

c) Put your code letter, *not* your name, on each page.

d) Return the examination with a signed statement of honor pledge on a page separate from your answers.

e) You are required to answer *only what is asked* in the questions and *not all you know* about the topics.

Question 1

Let the joint density function of two continuous random variables X and Y be

$$f_{X,Y}(x,y) = (y + x\theta)/\theta^2 \; ; \quad 0<x<1 \; , \quad 0<y<\theta \; .$$

a) Show that the marginal distribution of X is

$$f_X(x) = \tfrac{1}{2} + x \; , \quad 0<x<1 \; .$$

b) Show that the conditional distribution of Y given $X = x$ is

$$f_Y(y|x) = 2(y + x\theta)/\theta^2(1 + 2x) \; , \quad 0<y<\theta \; .$$

c) $E(Y|X = x)$ is often referred to as the "regression equation" of Y on x. For this problem, show that

$$E(Y|X = x) = \theta z \; ,$$

where $z = (2 + 3x)/3(1 + 2x)$.

(#7 Continued)

Question 2

Suppose that a survey is done to estimate a rate (P^*) which can be expressed as a constant (K) times a proportion (P). Assuming simple random sampling (SRS), the appropriate estimator of P^* is $p^* = Kp$, where p is the corresponding sample proportion. Ignoring the finite population correction, verify the following properties of p^* for an element sample of size n from a population with N elements:

a) True Variance: $\mathrm{Var}(p^*) = \dfrac{NP^*(K-p^*)}{n(N-1)}$

b) Coefficient of Variation: $\mathrm{CV}(p^*) = \dfrac{N(K-P^*)}{n(N-1)P^*}$

c) Assuming that $K = 10^3$, $P^* = 20$, and $N = 10^6$ (so that the finite population correction can be ignored), how large a sample would be needed in the survey yo yield a standard error or 5 for the estimator p^*?

d) Suppose that it is desired to estimate the parameter θ using n pairs of data points (x_i, Y_i), $i = 1,2,\ldots,n$. A reasonable criterion for choosing an estimator $\hat{\theta}$ of θ is to choose $\hat{\theta}$ to minimize the quantity

$$\sum_{i=1}^{n} (Y_i - \theta z_i)^2 \quad .$$

Show that the choice for $\hat{\theta}$ which minimizes the above expression is

$$\hat{\theta} = \sum_{i=1}^{n} z_i Y_i \Big/ \sum_{i=1}^{n} z_i^2 \quad .$$

e) For the $n = 3$ pairs of (x, Y) data points $(-1, 1)$, $(0, 0)$ and $(1, 3)$, what is the value of $\hat{\theta}$?

Question 3

Let the continuous random variables X and Y have joint distribution

$$f_{X,Y}(x,y) = (4xy)^{-\frac{1}{2}}, \quad 0<x<y<1 \quad .$$

a) Find the joint density of the random variables

$$U = \sqrt{X/Y} \quad \text{and} \quad V = \sqrt{Y} \quad .$$

b) Demonstrate whether or not U and V are independent random variables.

c) Find $E(Y|X = x)$.

Question 4.

To assess the harmful effects of certain environmental contaminants (e.g., arsenic or lead), researchers often adopt an experimental strategy which involves injecting a number of pregnant female mice with different doses of the substance in question and then observing the effects on the fetuses in each litter. One possible statistical model for the situation where a continuous response (e.g., fetal birth weight) is recorded for each fetus is

$$Y_{ij} = \eta_i + \varepsilon_{ij} \quad ,$$

where Y_{ij} is the observed response associated with the j-th fetus in the i-th litter (or, equivalently, dose level). Consider the following two different sets of assumptions concerning the above model:

(#4 Continued)

ASSUMPTION SET I:

(i) $\eta_i \sim N(\mu_i, \sigma_1^2)$

(ii) $\varepsilon_{ij} \sim N(0, \sigma_2^2)$

(iii) the $\{\eta_i\}$ and $\{\varepsilon_{ij}\}$ are all mutually independent random variables.

ASSUMPTION SET II:

(i) η_i is a fixed, unknown constant representing the "effect" of the i-th dose level.

(ii) $\varepsilon_{ij} \sim N(0, \sigma^2)$

(iii) for $j \neq j'$,

$$\text{cov}\ (\varepsilon_{ij}, \varepsilon_{i'j'}) = \begin{cases} \Theta & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases}$$

(i.e., observations in the same litter are correlated).

a) What is $E(Y_{ij})$ under Assumption Set I and under Assumption Set II?

b) What is $\text{Var}\ (Y_{ij})$ under Assumption Set I and under Assumption Set II?

c) What is the correlation between Y_{ij} and $Y_{ij'}$ for $j \neq j'$ (i.e., between two observations in the same litter) under Assumption Set I and under Assumption Set II?

d) Under Assumption Set I, show that Y_{ij} and $Y_{ij'}$ for $j \neq j'$ are independent conditional on η_i being fixed.

NOTE: $\text{cov}\ (Y_{ij}, Y_{ij'} | \eta_i) = E[Y_{ij}Y_{ij'} | \eta_i] - E(Y_{ij} | \eta_i)E(Y_{ij'} | \eta_i)$

$$= \eta_i^2 - \eta_i\eta_i = 0 \quad .$$

BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(April 14, 1985)

INSTRUCTIONS:

a) This is an *open book* examination.

b) M.P.H. students are to answer *any two* questions during the three-hour period (2 PM - 5 PM). M.S. students are to answer *three* questions -- time period 1 PM - 5 PM.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, *not* your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

f) You are required to answer *only what is asked* in the questions and *not all you know* about the topics.

Question 1.

Two faculty members independently graded the responses to an exam question answered by seven students. Each grader assigned a score between 0 and 25 to each response. Professor One stored his grades in a SAS dataset in a catalogued SAS data library stored on online disk at TUCC. The PROC CONTENTS output for this dataset is attached. Professor Two stored his grades on a scrap of paper, reproduced below. Write a program, including all necessary JCL and SAS statements to do the following:

1) Produce a scatter plot of Professor One's score versus Professor Two's score, using the student's ID letter as the plotting symbol. Title and label the plot appropriately.

2) Produce a report listing each student whose two scores differ by five or more points. The listing should include the students ID, the two scores, and the difference between the scores.

NOTES: In your JCL use the account code UNC.B.E1234, a programmer name of SMITH, and a password of SECRET. You may assume that each student has exactly one record in each "file", and that no values are missing in either "file".

CONTENTS OF SAS DATA SET ALPHA.EXAM

TRACKS USED=2 SUBEXTENTS=1 OBSERVATIONS=7 CREATED BY CS JOB ICOK ON CPUID 23-3081-021232 AT 16:23 WEDNESDAY, MARCH 20, 1985
BY SAS RELEASE 82.4 DSNAME=UNC.E.E522Y.ONE.EXAMLIB BLKSIZE=19062 LRECL=13 OBSERVATIONS PER TRACK=1466 GENERATED BY DATA

ALPHABETIC LIST OF VARIABLES

#	VARIABLE	TYPE	LENGTH	POSITION	FORMAT	INFORMAT	LABEL
2	GRADE	NUM	8	5			EXAM SCORE FOR PROF. ONE
1	ID	CHAR	1	4			STUDENT IDENTIFICATION

Question 1, Grader Two

Student	Score
A	9
F	12
S	18
B	6
C	22
E	24
G	14

Question 2.

Provide the p-value for the following comparison of two drugs, where the new drug is expected to be at least as effective as the old one.

	Outcome		
	Improved	Not Improved	Total
New Drug	6	6	12
Old Drug	1	7	8
Total	7	13	20

a) Use Pearson's chi-square test, with and without Yates correction.

b) Use Fisher's Exact test.

c) Use the normal approximation to Fisher's Exact test.

d) Summarize briefly these results for the comparison of the two drugs.

Question 3.

a) What is the difference between regression with an intercept and regression through the origin?

b) Perform an appropriate regression analysis for the following data, X is carrier and Y the dependent variable.

X	Y
3	7
4	7
6	12
8	17
7	13

$\Sigma X = 28 \qquad \Sigma Y = 56 \qquad \Sigma XY = 348$

$\Sigma X^2 = 174 \qquad \Sigma Y^2 = 700$

c) Repeat the regression analysis of Part b using centered data.

d) Is there a difference between regression through the origin and regression using centered data? Explain.

e) Suppose a sixth data point for Part b was taken and it was $X = 0$, $Y = 1$. Which of the procedures you have considered (regression with intercept, regression through origin, regression with centered data) become inapplicable? Explain.

Question 4.

Consider a *population* composed of the following $N = 30$ elements:

$$Y_1 = 1; \quad Y_2 = 2, \ldots, Y_{30} = 30 \quad .$$

a) Compute the true variance of the mean of a simple random sample of size $n = 6$.

b) If the population were divided into $H = 6$ strata as follows:

$$(Y_1, Y_2, Y_3, Y_4, Y_5); \ (Y_6, Y_7, Y_8, Y_9, Y_{10}); \ \ldots; \ (Y_{26}, Y_{27}, Y_{28}, Y_{29}, Y_{30}),$$

what is the true variance of the estimate of the population mean for a stratified simple random sample of $n = 6$ chosen by selecting $n_h = 1$ element in each stratum?

c) Verify whether or not the design in part (b) is epsem.

HINT: Recall that $\sum_{i=1}^{N} i = N(N+1)/2$ and $\sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6$.

WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(January 18, 1986; 9:30 AM to 12:30 PM)

Question 1

After injecting a unit amount of a certain radioactive substance into the blood stream of an individual at time $t = 0$, suppose that the ratio $Y = P/(1 - P)$ of the proportion P of this substance remaining in the blood stream at time $t(>0)$ relative to the proportion $(1 - P)$ having been eliminated from the blood stream is a random variable with conditional CDF

$$F_Y(y|T = t) = (1 - e^{-ty}) \ , \ 0 < y < +\infty,$$

where the time variable T has the uniform distribution over the interval $(0, 1)$, namely,

$$F_T(t) = t \ , \ 0 < t < 1.$$

a) Find the unconditional CDF $F_Y(y)$, of Y directly.

b) Find $f_Y(y)$, the marginal density function of Y.

c) Find $E(Y)$ and $V(Y)$ directly.

d) Find $E(Y)$ and $V(Y)$ by appropriately unconditionalizing $E(Y|T = t)$ and $V(Y|T = t)$ over the stated distribution of T, and then compare [illegible] answers to those obtained in part (c).

e) [illegible] (Y, T).

f) [illegible] the relationship $Y = P/(1 - P)$, find that function $g(p)$ for which

$$F_p(p) = F_Y[g(p)] \ ,$$

and then use this result to find $f_p(p)$.

Question 2

According to a certain genetic theory, the expected proportions of three genotypes in offspring from certain crosses of rats are as follows:

Genotype 1: $1/(3 + \theta)$;

Genotype 2: $2/(3 + \theta)$;

Genotype 3: $\theta/(3 + \theta)$;

here, $\theta (\geq 0)$ is an unknown parameter. Suppose that n such offspring are selected at random and tested appropriately. Let the random variable X_i denote the number of genotype i observed, i = 1, 2, 3; thus, $\sum_{i=1}^{3} X_i = n$.

a) Using the fact that the joint distribution of X_1, X_2, and X_3 is multinomial, show that X_3 is a sufficient statistic for θ.

b) It is of interest to test H_0: $\theta = 1$ (the value expected under standard Mendelian theory) versus H_1: $\theta \neq 1$. Find an explicit expression for the likelihood ratio statistic $\hat{\lambda}$ for testing H_0 versus H_1. For large n, how would you use $\hat{\lambda}$ to test H_0 versus H_1?

c) By expressing $\hat{\lambda}$ in an appropriate form, show that the P-value for a test of H_0: $\theta = 1$ versus H_1: $\theta \neq 1$ can be calculated using the binomial distribution.

Question

A screening program for hypertensive patients in a certain community requires that a potential patient have n independent readings of his or her blood pressure taken at each of k separate clinic visits. The reason for requiring a total of $k \cdot n$ blood pressure readings on each subject is that blood pressure readings are known to fluctuate considerably, both between and within visits. A model used to account for these two sources of variability is as follows. If Y_{ij} denotes the observed blood pressure reading for a randomly selected individual for the j-th of the n readings taken at the i-th visit, then

$$Y_{ij} = \theta + \varepsilon_i + \eta_{ij} ,$$

where θ is a fixed unknown constant, ε_i $N(0, \sigma_B^2)$, η_{ij} $N(0, \sigma_W^2)$, and ε_i and η_{ij} are independent, $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$.

a) Find $E(Y_{ij})$, and discuss its meaning with regard to the hypertension screening program described above.

b) Find $V(Y_{ij})$.

c) If $\bar{Y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$, find $V(\bar{Y}_{i.})$.

d) Find $E(e^{tY_{ij}})$, and then use the result to infer the form of the distribution of Y_{ij}.

Question 4

Suppose that associated with every member of some population of N elements are two measures, a y-variable and an x-variable. When a simple random sample of size n is chosen, the sample means, $\bar{y}_o$ and $\bar{x}_o$, may be used to estimate the corresponding population means, $\bar{Y}$ and $\bar{X}$, respectively.

a) The covariance between $\bar{y}_o$ and $\bar{x}_o$ is

$$\text{Cov}\ (\bar{y}_o, \bar{x}_o) = (1-f)\ S_{yx}/n,$$

where $f = n/N$ and S_{yx} is another covariance. Give a clear intuitive explanation of the conceptual difference between $\text{Cov}(\bar{y}_o, \bar{x}_o)$ and S_{yx}.

b) Note that $r = \bar{y}_o/\bar{x}_o$ may be used to estimate $R = \bar{Y}/\bar{X}$ when n is sufficiently large. Decompose the mean square error of r, $\text{MSE}(r) = E(r - R)^2$, into bias and variance components.

c) One rationale for obtaining a variance expression for r is to note that

$$r - R = (\bar{y}_o - R\bar{x}_o)/\bar{x}_o \doteq (\bar{y}_o - R\bar{x}_o)/\bar{X}$$

Using this rationale and stating any additional assumptions, develop an expression for Var(r).

d) Assuming that the two measures are uncorrelated in the population and that r is unbiased, show that the coefficient of variation for r is approximately

$$\text{CV}(r) = [\{\text{CV}(\bar{y}_o)\}^2 + \{\text{CV}\ (\bar{x}_o)\}^2]^{1/2}$$

MS WRITTEN EXAMINATION IN BIOSTATISTICS

PART II: APPLICATIONS

(January 19, 1986; 1:00 PM to 5:00 PM)

INSTRUCTIONS

a) This is an open-book examination.

b) The time limit is four hours.

c) Answer any three of the four questions which follow.

d) Put the answers to different questions on separate sheets of paper.

e) Since your papers will be xeroxed for back-up purposes, use a pen/pencil and paper combination which will reproduce clearly.

f) Put your code letter, not your name, on each question.

g) Return the examination with a signed statement of the honor pledge on a page separate from your answers. Do not identify your code letter on the honor pledge; use your name.

Question 1

Four vaccines were treated with six evenly spaced levels of an additive that was supposed to increase the antibody response. The data, calculations, and summary statistics are provided in Table 1 for an analysis of variance.

a) Perform an analysis of variance, reporting on the possibility that any dose response pattern may depend upon the vaccine.

b) Presuming a common dose response pattern across vaccines, examine the significance and adequacy of a linear trend on the amount of additive.

c) Calculate 95% confidence limits for the true average antibody response, presuming a true linear trend, at the mean amount of additive.

Table 1

INFLUENZA ANTIBODY RESPONSES TO 4 VACCINES WITH 6 DIFFERENT AMOUNTS OF ADDITIVE

Vaccine	*Amount of Additive* I	II	III	IV	V	VI	*Row Sum* (R_i)	*Row Mean* ($\bar{x}_{i..}$)
	5	2	3	7	3	7		
A	6	4	3	4	8	8	87	4.83
	5	4	6	3	6	3		
	3	3	5	2	6	4		
B	2	6	7	7	3	7	82	4.56
	4	3	6	4	4	6		
	5	5	6	5	9	3		
C	2	3	7	6	7	6	95	5.28
	2	6	4	7	4	8		
	2	4	2	7	5	5		
D	4	2	2	2	6	2	59	3.28
	2	3	2	3	2	4		
Column Sum (C_j)	42	45	53	57	63	63	323 (T)	
Column Mean ($\bar{x}_{.j.}$)	3.50	3.75	4.42	4.75	5.25	5.25		4.49 ($\bar{x}$)

CELL TOTALS (T_{ij})

Vaccine	*Amount* I	II	III	IV	V	VI
A	16	10	12	14	17	18
B	9	12	18	13	13	17
C	9	14	17	18	20	17
D	8	9	6	12	13	11

$\Sigma\Sigma\Sigma x_{ijk}^2 =$

$\Sigma c_j^2 =$

$\Sigma\Sigma T_{ij}^2 =$

$\Sigma R_i 2 =$

Question [illegible]

[illegible] children, the hypothesis is that the death of a prev[illegible] [illegible] the behavior of surviving children.

[illegible] control data were collected.

Birth order	problem [illegible]	Number of mothers with Losses	No Losses	Total
1	Problem child	20	82	102
	Control	10	54	64
		30	136	166
2-4	Problem child	26	41	67
	Control	16	30	46
		42	71	113
5+	Problem child	27	22	49
	Control	14	23	37
		41	45	86

a) State and test an appropriate hypothesis that controls for the effect of birth order. Use $\alpha = 0.005$.

b) Discuss the examination of the idea that as the number of sibling deaths increase, the problem child becomes more common. What re-tabulation of these data might be helpful in this regard?

Question [illegible]

Consider the following times to remission for acute leukemia patients.

[illegible]	1-2	3-4	5-6	7-8	9-10	11-12
[illegible]	1	3	2	4	0	4

[illegible]	15-16	17-18	19-20	21+
	1	1	1	2

The two [illegible] observations in the right tail were 22 and 23 weeks.

a) Discuss very briefly candidate models for a summarization of these data. Parsimony in the number of required parameters should be a guiding thought. Select and defend a model.

b) Estimate the parameter(s) of your chosen model.

c) Examine the fit of this model to these data.

Question 4

An experiment was conducted by a private research corporation to investigate the toxic effects of three chemicals (I,II,III) used in the tire-manufacturing industry. In this experiment one-inch squares of skin on rats were treated with the chemicals and then scored from 0 to 10 depending on the degree of irritation. Three adjacent 1-inch squares were marked on the backs of 8 rats, and each of the three chemicals was applied to each rat.

a) Recognizing this as a randomized blocks design, what are the blocks and what are the treatments?

b) State the assumptions on which the analysis is based, and indicate how you could examine the validity of each assumption.

c) A partial ANOVA table is shown below. Fill in the missing entries.

Source	df	SS	MS	F
Treatments	___	25.0	___	___
Blocks	___	18.5	___	___
Error	___	___	___	___
Total	___	68.5		

d) Do the data provide sufficient evidence to indicate a significant difference in the toxic effects of the three chemicals?

e) The means of the responses to the three chemicals are shown below:

I	6.25
II	7.50
III	5.00

Given that the .05 critical value for any pairwise difference between means is ± 1.75, which of the pairs of chemicals give significantly different mean responses?

f) The critical value in part (e) is obtained using Tukey's procedure for multiple comparisons. Why is it necessary to use a procedure such as this?

g) What conclusions can you draw from this analysis?

[illegible] WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

[illegible] 12, 1986: 9:30-12:30PM

INSTRUCTIONS:

a) This is a closed book examination.

b) Answer any three questions during the three hour time period.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, not your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

QUESTION 1

We want to investigate whether drugs A and B taken orally have any long term carcinogenic side effects. The response variable will be the number of malignant tumors, treated as a continuous variable. 150 mice are divided into three groups of 50 mice each (Control, A and B), and fed the appropriate diets for two years.

There is reason to believe that the carcinogenic effect is influenced by caloric intake of the mice over the duration of the experiment. Although we are not interested in this "caloric effect" per se, we need to take it into account. Assume that by weighing leftover mouse chow, a fairly good estimate can be provided for caloric intake.

We want to test the hypothesis that there is no difference between mice on the three diets.

(1) What assumption needs to be made in order to use covariate analysis? Please state it in terms of the variables in the problem.

(2) Write down the model needed to test the assumption in (1). State the null hypothesis for this test in terms of the parameters of [illegible]. Define all your variables!

(3) [illegible] can not make the necessary assumption for covariate [illegible] different hypothesis could you test to distinguish [illegible] groups of mice? State the hypothesis in terms of [illegible] of your model. Give the F-statistic used to test the [illegible] in terms of sums of squares associated with the variables in [illegible] model. Include degrees of freedom.

(4) If the assumption in (1) is not rejected (so that covariate analysis can be used), write down a model that could be used to test whether or not there is a difference between the three diets. Define all variables and write the null hypothesis in terms of the parameters of your model.

(5) Give the F-statistic used to test the hypothesis in (4) in terms of sums of squares associated with variables in your model. Include the degrees of freedom for the numerator and denominator.

QUESTION 2

Consider the following population of 1,000 individuals who have been divided into the following strata:

h	W_h	S_h^2
1	0.1	64
2	0.4	100
3	0.5	25

where W_h is the proportion of the population falling in the hth stratum and S_h^2 is the stratum-specific element variance. A proportionate stratified simple random sample of individuals is to be chosen for a survey whose dual purpose is to estimate the population mean per individual ($\bar{Y}$) as well as the difference in means between the first and second strata (i.e., $D = \bar{Y}_1 - \bar{Y}_2$).

a. Briefly describe the steps you would follow in selecting the sample.

b. How large an overall sample (i.e., for all three strata combined) would be needed for the estimator (d) of D to have a standard error of 2?

c. Assuming per-unit interviewing costs to be equal among strata, what would be the relative sizes of the three stratum sampling rates ($f_n = n_h/N_h$) if optimum stratum allocation were used?

QUESTION 3

Let

$$f_Y(y) = \theta e^{-\theta y} \quad , \; y > 0 \; ,$$

be the density function for a continuous random variable Y. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from this continuous distribution.

a. What is the joint distribution (i.e., the likelihood function) of $Y_1, Y_2, \ldots, Y_n$?

b. Show that the maximum likelihood estimator of θ is

$$\hat{\theta} = (\bar{Y})^{-1} \; ,$$

where $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

c. Show that $\hat{\theta}$ is a sufficient statistic for θ.

d. Find an appropriate large sample 95% confidence interval for θ if $n = 100$ and $\bar{Y} = 4$.

QUESTION [illegible]

Resea[illegible] that the prevalence (Y) [illegible] byssinosis in workers in textil[illegible]ing plants is linearly related to the mean daily cotton dust [illegible] Under the assumption that zero cotton dust level implies zero prevalence, the regression equation relating the mean of Y to X is

$$E(Y) = \beta X.$$

a. Given the n pairs of data points (X_i, Y_i), $i = 1, 2, \ldots, n$, show that the least-squares estimator of β is

$$\hat{\beta} = \sum_{i=1}^{n} X_i Y_i \Big/ \sum_{i=1}^{n} X_i .$$

Now, henceforth assume that each Y_i is normally distributed with mean βX_i and variance σ^2, that the Y_i's are mutually independent, and that the X_i's are fixed constants (i.e., they are not random variables). With these assumptions, show that:

b. $E(\hat{\beta}) = \beta$.

c. $Var(\hat{\beta}) = \sigma^2 \Big/ \sum_{i=1}^{n} X_i^2$.

d. $\hat{\beta}$ is normally distributed.

e. Using the results in parts (b), (c), and (d), develop a test of $H_o: \beta = 0$ versus $H_a: \beta \neq 0$ under the assumption that σ^2 is known. In particular, give the appropriate test statistic and define the rejection region precisely so that Pr (Type I error) = 0.01.

SPECIAL MS WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

April 13, 1986: 9:30AM to 1:30PM

INSTRUCTIONS:

a) This is an open book examination.

b) Answer any three questions.

c) Put the answers to different questions on separate sets of papers.

d) Put your CODE LETTER, (not your name), on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

f) You are required to answer only what is asked in the questions, and not all you know about the topics.

QUESTION 1

a. (10 points: 1 pnt/part)
Briefly describe or explain (in one or two sentences) the following terms:

(1) OS
(2) TSO
(3) Track (on magnetic disk)
(4) Block
(5) Parity bit
(6) Field
(7) JCL
(8) EBCDIC
(9) Collating sequence
(10) Backup

b. (15 points: 3 pnts/part)
In a brief paragraph, possibly using diagrams or Figures, compare and contrast each group of terms:

(1) OS file, SAS data library, SAS dataset
(2) Logical record, physical record
(3) JCL step, SAS step
(4) DSN, ddname
(5) Program data vector, input buffer

[illegible]

[illegible] of 230 individuals was typed for the MN [illegible] results as follows:

[illegible]	Number of individuals
[illegible] (f_1)	[illegible]
[illegible] (f_2)	111
[illegible] (f_3)	[illegible]
Total (T)	230

The maximum likelihood estimates of the allele frequencies of M and N are given by

$$\hat{p} = \text{freq}(M) = \frac{2f_1 + f_2}{2T}$$

$$\hat{q} = \text{freq}(N) = 1 - \hat{p}$$

$$\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{2T} = \text{Var}(\hat{q})$$

At equilibrium, we expect the genotypic frequencies to be in the proportion

$$p^2 : 2pq : q^2 \text{ , MM:MN:NN.}$$

Is there evidence that the system deviates from equilibrium conditions?

[illegible] were given in a course. The first exam had average [illegible] deviation 1[illegible]. The second exam had average score [illegible] 15. The correlation between the two scores [illegible] scored 60 points on the first exam, what score [illegible] student to have on the second exam? How good is [illegible] would be your predicted value when the correlation [illegible] is equal to (a) [illegible] and (b) -0.2?

QUESTION 4

Consider the following regression model used to analyze data from a dental examination survey.

$$\text{Oral health hygiene score} = \text{Mean level} + \text{Age effect} + \text{Sex effect} + \text{Education effect} + \text{Income effect} + \text{Error}$$

where age, education, and income are regarded as continuous variables and sex is categorical. The following ANOVA table summary was transcribed from a PROC REG printout when the above regression model was fitted to the data.

Source	DF	SS	MS	F
Model	4	22.613	5.653	7.733
Error	256	187.136	0.731	
Corrected total	260	209.749		

Source	DF	Type I ss	F	DF	Type II ss	F
Age	1	8.486	11.609	1	6.987	9.558
Sex	1	7.371	10.083	1	7.165	9.802
Education	1	6.555	8.967	1	0.676	0.925
Income	1	0.201	0.275	1	0.201	0.275

A table of F-values is attached to facilitate their interpretation.

Now answer the following questions:

(i) Define sequential F-tests and partial F-tests as used in regression modelling.

(ii) Identify the numerical results for sequential and partial F-tests from the ANOVA summary presented.

(iii) Why do the Type I ss add to the model ss but the Type II ss not add to the model ss?

(iv) What is the sample size used in the regression analysis of the oral hygiene scores?

(v) Should income be included in the regression model? Explain your answer. If you feel you need further information to make the assessment, explain what information you need.

(vi) If you were to develop a regression model for predicting oral health hygiene scores and your potential explanatory variables were age, sex, race, education, income, and number of visits to the dentist during the last year, how would you proceed. List the methodological steps you would undertake.

(vii) If your objective was how to assess the impact of frequency of visits to the dentist on oral health hygiene scores rather than predict oral health hygiene, would you proceed any differently to your answer in part (vi).

WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

[illegible] 24, 1987 10:00 AM to 1:00 PM

INSTRUCTIONS:

a) This is a closed book examination.

b) Students are to answer three questions.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, not your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

Question 1

For a certain multiple choice question with m possible choices (only one of which is correct), suppose that n randomly chosen students of equal ability attempt the question. Let θ equal the probability that a student actually knows the right answer to the question. Then, $(1 - \theta)$ is the probability that a student does not really know the answer to the question (i.e., the student is guessing); in this case, the probability that a student answers the question correctly, given that he or she is guessing, is m^{-1}.

a) Prove [illegible] that the probability π that a student answers the [illegible] is equal to

$$\pi = [1 + \theta(m - 1)]/m.$$

b) Let [illegible] variable X denote the number of students out of n that answer the question correctly. Find the maximum likelihood estimator $\hat{\theta}$ of θ.

c) Show that $\hat{\theta}$ is an unbiased estimator of θ, and also find the variance of $\hat{\theta}$.

d) For the special case $m = 2$ and $\theta = \frac{1}{2}$, show that $\hat{\theta}$ achieves the Cramér-Rao lower bound for the variance of any unbiased estimator of θ.

Question 2

An investigator is studying the response (sleeping time) to a drug (tranquilizer) at three equally spaced dose levels. The drug is given to n different animals at each level.

a) Set up an ANOVA table and interpret the sources of variation and the possible tests of hypotheses.

b) What are the underlying assumptions of the procedure?

c) How would the sources of variation and degrees of freedom in the ANOVA table change if:

1. The same animals were used at each dose level
2. Only two dose levels were used
3. Four dose levels were used
4. Three different drugs were used instead of three levels of the same drug.

[illegible] health awareness program to get people to consume [illegible] cholesterol foods is being evaluated by the following means. Two cities, each with 50,000 residents and located in the same state, are chosen for the evaluation study. A baseline measure of serum cholesterol is first measured simutaneously on a (without replacement) simple random sample of n residents separately chosen in each city. One of the cities (the <u>intervention</u> city) is then subjected to the new awareness program for a period of four years, while the other (the <u>control</u> city) uses existing awareness strategies during the same period. At the end of the four year period a followup measure of cholesterol is obtained from the same samples in the two cities. We shall assume for simplicity: (1) that the populations in the two cities do not change in composition during the study, and (2) that there is no attrition in the samples between baseline and followup.

The object of analysis will be to estimate the difference in the baseline-to-followup mean differences in the two cities, $\delta_I - \delta_C$, where

$$\delta_I = (\bar{Y}_f - \bar{Y}_b) \quad \text{and} \quad \delta_C = (\bar{X}_f - \bar{X}_b)$$

and the subscripts "f" and "b" denote followup and baseline, respectively. The baseline-to-followup <u>sample</u> mean difference,

$$\hat{\delta}_I - \hat{\delta}_C = (\bar{y}_f - \bar{y}_b) - (\bar{x}_f - \bar{x}_b),$$

is used to estimate $\delta_I - \delta_C$, where $\bar{y}_f$, $\bar{y}_b$, $\bar{x}_f$, and $\bar{x}_b$ are sample means corresponding to the four means in δ_I and δ_C.

[illegible]w that $\hat{\delta}_I - \hat{\delta}_C$ is an unbiased estimator of $\delta_I - \delta_C$.

[illegible]ssuming the same population element variance for [illegible]holesterol (S^2) in each city during baseline and [illegible]followup, show that

$$Var(\hat{\delta}_I - \hat{\delta}_C) = [4 - 2(\rho_y + \rho_x)](1-f)S^2/n,$$

where ρ_y is the correlation between the baseline and followup measure of cholesterol in the intervention city, ρ_x is the comparable correlation for the control city, and f is the sampling rate in both cities.

(#3 Continued)

c. Derive the formula you would use to determine the sample size (n) needed to achieve a prescribed coefficient of variation for the baseline-to-followup sample mean difference.

d. Roughly estimate how much the sample sizes determined from the formula in Part (c) would be affected if an unstratified two stage cluster sampling design (with blocks as PSUs) were used instead of simple random sampling. Assume that the average sample size per PSU in the cluster designs would be 10 and that the intraclass correlation for cholesterol in both communities would be 0.03 for these clusters.

Question 4

According to a certain genetic theory, the expected proportions of three genotypes in offspring from certain crosses of rats are as follows:

Genotype 1: $1/(3 + \theta)$;

Genotype 2: $2/(3 + \theta)$;

Genotype 3: $\theta/(3 + \theta)$;

here, $\theta (\geq 0)$ is an unknown parameter. Suppose that n such offspring are selected at random and tested appropriately. Let the random variable X_i denote the number of genotype i observed, i = 1, 2, 3; thus, $\sum_{i=1}^{3} X_i = n$.

a) Using the fact that the joint distribution of X_1, X_2 and X_3 is multinomial, show that X_3 is a sufficient statistic for θ.

b) It is of interest to test H_0: $\theta = 1$ (the value expected under standard Mendelian theory) versus H_1: $\theta \neq 1$. Find an explicit expression for the likelihood ratio statistic $\hat{\lambda}$ for testing H_0 versus H_1. For large n, how would you use $\hat{\lambda}$ to test H_0 versus H_1?

c) By expressing $\hat{\lambda}$ in an appropriate form, show that the P-value for a test of H_0: $\theta = 1$ versus H_1: $\theta \neq 1$ can be calculated using the binomial distribution.

[illegible]

[illegible]

[illegible]

The article below appeared in the NEW YORK TIMES for 30 June 1951. Calculate the statistical significance of the results in the fourth and fifth paragraphs, and comment on the conclusions in the last paragraph. Note: Willard Parker Hospital was, until it closed, a famous center for treating contagious diseases.

TRAVEL HELD RAISING POLIO FATALITY RATE. A study indicating that transportation over long distances greatly increases the death rate among victims of infantile paralysis in the acute stage is described in the new issue of the JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION.

The results were reported by Dr. M. Bernard Brahdy of Mount Vernon, New York, and Dr. Selig H. Katz of New York, both of whom are associated with the Willard Parker Hospital in Manhattan.

The report is based on a study of the records of 493 polio victims admitted to the hospital. Of these, 380 were local patients who travelled an average of seven miles to the hospital. The other [illegible] were transported an average of eighty-five miles. The study covered the [illegible] epidemic in the summer and fall of 1949.

There [illegible] deaths among the 380 local patients, a fatality rate [illegible]. In contrast, there were eighteen deaths [illegible] transported patients, a death rate of 16 percent.

[illegible] of the deaths among the local group occurred [illegible] after admission to the hospital, fully [illegible] miles among the transported group occurred [illegible] average duration of illness before admission was [illegible] both groups -- three and a fraction days.

It seems, the report said, "that the greater mortality in the transported group, occurring shortly after admission to the hospital is a manifestation of the effect of long transportation during the acute stage of illness."

Question 2

This problem is based on a study in which each of 280 subjects was classified according to sex (female, male), participation in a particular health program (yes,no), and attitude toward a particular health policy (favorable, unfavorable). The frequency distribution of these subjects with respect to sex x participation x attitude is as follows:

		Attitude		
Sex	Participation	Unfavorable	Favorable	Total
Female	No	30	10	40
Female	Yes	51	49	100
	TOTAL	81	59	140
Male	No	25	75	100
Male	Yes	4	36	40
	TOTAL	29	111	140

a. Evaluate the extent of association between attitude and participation for all subjects

b. Evaluate the extent of association between attitude and participation for males. Also do this for females.

c. Determine the log {odds ratio} measure of association between attitude and participation for males. Also do this for females. Evaluate the similarity of the log {odds ratio} for males to that for females (Hint: The variance of the log {odds ratio} for a (2x2) table is approximately $\sum_{i=1}^{2} \sum_{j=1}^{2} (1/n_{ij})$ where n_{ij} is the frequency in the i=th row and j-th column).

d. Evaluate the association between attitude and participation for all subjects in a way which adjusts for sex.

e. Evaluate the association between sex and attitude for all subjects and between sex and participation for all subjects. Discuss these results in connection with the nature of conclusions for (a), (b), (c), and (d).

Exercise are from a study undertaken by [illegible] School of Dentistry. They pertain to the relation [illegible] cross-sectional area.

Subject	1	2	3	4	5	6	7	8	9	10
Area	0.10	0.21	0.24	0.27	0.28	0.31	0.35	0.38	0.39	0.39
Flow	168	277	352	463	435	601	505	640	558	702

a. Determine the least squares estimates for the intercept and slope of a linear relationship. Determine a 0.95 confidence interval for the slope.

b. Assess whether the inclusion of a quadratic term is significant at the $\alpha=0.05$ level.

c. Determine a 0.95 confidence interval for the expected flow when area=0.30? when area=0.40? Use the model in (a).

d. Determine a 0.95 prediction interval for a future observation at area=0.30? at area=0.40? Use the model in (a).

e. For the model in (b), test the hypothesis that the coefficients of area and (area)^2 are 0.

Question 4

Harbin et al. (1987) tested simple reaction time (in milliseconds) of young (age 18-28 years) and old (age 60-86 years) men. Median reaction time for 30 trials for each subject was recorded. The scores were found to not be detectably non-Gaussian, within group. Subjects were randomly assigned to either 0 ppm carbon monoxide (CO) exposure or 100 ppm CO exposure.

Using the output from the attached program fragment, conduct an appropriate ANOVA.

a.) Provide a source table, including df, SS, and F.

b.) Evaluate the significance (use α = .01) of tests in the table. Report your interpretation of the results.

c.) Conduct and report any appropriate stepdown tests as indicated by the results in b.). Report appropriate stepdown tests, means, differences to be tested. Use a pair-wire differences approach, for Bonferroni corrected t-tests.

d) Specify

i.) appropriate dummy variable codings
ii.) appropriate full rank model using the dummy variables

to use a regression program to conduct the ANOVA.

e) State each null hypothesis tested in the ANOVA table in part b.) in terms of the regression coefficients in d.).

f) The data listed are a balanced subset of 53 observations (14 young 0 ppm, 17 young 100 ppm, 13 old 0 ppm, 9 old 100 ppm).

Explain in one or two sentences one important statistical advantage of using a regression program rather than a hand calculator for such unbalanced data.

[illegible]TION TIME [illegible]

	GROUP	CO_LEVEL	[illegible]	[illegible]
	[illegible]	0PPM	1	391.5
	[illegible]	0PPM	2	471.0
	[illegible]	0PPM	3	492.0
	[illegible]	0PPM	4	376.0
	[illegible]NG	0PPM	5	372.5
	[illegible]NG	0PPM	6	307.0
	[illegible]UNG	0PPM	7	501.0
[illegible]	YOUNG	0PPM	8	466.0
9	YOUNG	0PPM	9	375.0
10	YOUNG	100PPM	1	472.5
11	YOUNG	100PPM	2	354.0
12	YOUNG	100PPM	3	350.0
13	YOUNG	100PPM	4	486.0
14	YOUNG	100PPM	5	402.0
15	YOUNG	100PPM	6	347.0
16	YOUNG	100PPM	7	320.0
17	YOUNG	100PPM	8	446.0
18	YOUNG	100PPM	9	410.0
19	OLD	0PPM	1	542.0
20	OLD	0PPM	2	599.0
21	OLD	0PPM	3	562.0
22	OLD	0PPM	4	586.0
23	OLD	0PPM	5	674.0
24	OLD	0PPM	6	762.0
25	OLD	0PPM	7	697.0
26	OLD	0PPM	8	583.0
27	OLD	0PPM	9	533.5
28	OLD	100PPM	1	523.5
29	OLD	100PPM	2	770.0
30	OLD	100PPM	3	712.0
31	OLD	100PPM	4	653.0
32	OLD	100PPM	5	699.5
33	OLD	100PPM	6	561.0
34	OLD	100PPM	7	751.0
35	OLD	100PPM	8	520.5
36	OLD	100PPM	9	523.0

[illegible]TA=ONE;

[illegible] USED 0.10 SECONDS AND 792K AND PRINTED PAGE [illegible]

[illegible] MEAN USS CSS SUM DATA=ONE;
[illegible]OUP CO_LEVEL NOTSORTED ;
[illegible]TIME ;

[illegible] THE PROCEDURE MEANS USED 0.10 SECONDS AND 808K AND PRINTED PAGE [illegible]

REACTION TIME DATA 2

---------- AGE GROUP, YOUNG OR OLD=YOUNG CO_LEVEL=0PPM ----------

VARIABLE	LABEL	N	MEAN	UNCORRECTED SS	CORRECTED SS
REACTIME	MEDIAN REACTION TIME, MS	9	428.0000000	1768840.50	120184.500

---------- AGE GROUP, YOUNG OR OLD=YOUNG CO_LEVEL=0PPM ----------

VARIABLE	SUM
REACTIME	3852.00000000

---------- AGE GROUP, YOUNG OR OLD=YOUNG CO_LEVEL=100PPM ----------

VARIABLE	LABEL	N	MEAN	UNCORRECTED SS	CORRECTED SS
REACTIME	MEDIAN REACTION TIME, MS	9	398.6111111	1458697.25	28679.8889

---------- AGE GROUP, YOUNG OR OLD=YOUNG CO_LEVEL=100PPM ----------

VARIABLE	SUM
REACTIME	3587.50000000

---------- AGE GROUP, YOUNG OR OLD=OLD CO_LEVEL=0PPM ----------

VARIABLE	LABEL	N	MEAN	UNCORRECTED SS	CORRECTED SS
REACTIME	MEDIAN REACTION TIME, MS	9	615.3888889	3457045.25	48713.8889

---------- AGE GROUP, YOUNG OR OLD=OLD CO_LEVEL=0PPM ----------

VARIABLE	SUM
REACTIME	5538.50000000

---------- AGE GROUP, YOUNG OR OLD=OLD CO_LEVEL=100PPM ----------

VARIABLE	LABEL	N	MEAN	UNCORRECTED SS	CORRECTED SS
REACTIME	MEDIAN REACTION TIME, MS	9	634.8333333	3712776.75	85656.5000

---------- AGE GROUP, YOUNG OR OLD=OLD CO_LEVEL=100PPM ----------

VARIABLE	SUM
REACTIME	5713.50000000

[illegible]

[illegible]

[illegible]

- This is a closed book examination.
- Students are to answer three questions of which [illegible] more than 2 should be from Group A - time period [illegible] AM - 1 PM.
- Put the answers to different questions on separate sets of papers.
- Put your code letter, not your name, on each page.
- Return the examination with a signed statement of honor pledge on a page separate from your answers.
- You are required to answer only what is asked in the questions and not all you know about the topics.

[illegible]

[illegible]

[illegible] be a random sample from the density [illegible]

[illegible] and [illegible].

[illegible] density function.

[illegible] using the method of moments.

[illegible] joint distribution of $X_1, X_2, \ldots, X_n$.

[illegible] maximum likelihood estimator of [illegible].

Question 1.

Consider the linear model $Y = \mu(x_1, x_2) + \varepsilon$ where $\mu(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$. Suppose the experimental data is as follows:

x_1	x_2	y
-2	-1	-13
-1	-1	-12
0	0	1
1	1	10
2	1	15

(a) Find the design matrix X.
(b) Find X'X.
(c) Find $(X'X)^{-1}$.
(d) Find the least square estimates of β_1 and β_2.

(e) Determine s, where s^2 is the usual unbiased estimate of σ^2.

(f) Find the two-sided 95% confidence interval for β_1 and β_2.

(g) Find the two-sided 95% confidence interval for $\beta_1 + \beta_2$ and $\beta_1 - \beta_2$.

(h) Determine approximate p-values for two-sided tests of each of the following hypotheses.

$$H: \beta_1 = 0; \quad H: \beta_2 = 0; \quad H: \beta_1 = \beta_2.$$

(i) Calculate the F statistic and determine the proper numbers of degrees of freedom for the usual F test of the hypothesis H: $\beta_1 = 0$ and $\beta_2 = 0$.

(j) Determine a two-sided 95% confidence interval for $\mu(2,3)$ and a two-sided 95% prediction interval for a new respnse Y corresponding to $x_1 = 3$ and $x_2 = 2$.

[illegible] $= 1, 2, \ldots,$ [illegible]

[illegible] there is a fixed probability p that [illegible] [illegible] the continuous [illegible] [illegible] to be mutually [illegible] [illegible] the total number [illegible] during the [illegible]

[illegible] that [illegible] what is the <u>[illegible]</u> [illegible] particular cell [illegible]

[illegible] part (a), show that the <u>unconditional</u> <u>distribution</u> [illegible] is Poisson with mean λp.

[illegible] that a particular cell will [illegible] at least k structural changes. Given the [illegible] what is the probability that a particular cell will [illegible]?

[illegible]: You do <u>not</u> need to be able to work parts (a) and (b) in order [illegible] this question.]

GROUP B

Question 4.

A scientist at the National Institute of Environmental Health Sciences is studying the toxic effects of a certain chemical by injecting a group of pregnant rats with the chemical and then observing the number of offspring with abnormalities in each litter.

Suppose that p is the probability that a newborn rat has an abnormality. Further, for the i-th out of N litters each of size two, let the random variable X_{ij} be defined so that

$X_{ij} = 1$ if the j-th newborn rat has an abnormality,

$X_{ij} = 0$ if the j-th newborn rat does not have an abnormality, $j = 1, 2$.

Since the two newborn rats from the same litter would seem to have a "natural" relationship to one another, the following joint distribution for X_{i1} and X_{i2} is proposed to allow for correlation between the responses of rats in the same litter:

(X_{i1}, X_{i2})	$pr(X_{i1}, X_{i2})$
(1, 1)	$p^2 + \Theta$
(1, 0)	$p(1 - p) - \Theta$
(0, 1)	$p(1 - p) - \Theta$
(0, 0)	$(1 - p)^2 + \Theta$

Show that $E(X_{i1}) = E(X_{i2}) = p$, $V(X_{i1}) = V(X_{i2}) = p(1 - p)$, and that $cov(X_{i1}, X_{i2}) = \Theta$.

What is the probability distribution of the random variable $Y_i = (X_{i1} + X_{i2})$?

MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

8 May 1987

INSTRUCTIONS:

a) This is an **open book** examination.

b) M.P.H. students are to answer **any two** questions during the three-hour period (10 AM - 1 PM). M.S. students are to answer **three** questions -- time period 10 AM - 2 PM.

c) Put the answers to different questions on separate sets of papers.

d) Put your code letter, **not** your name, on each page.

e) Return the examination with a signed statement of honor pledge on a page separate from your answers.

Editorial note: Question 1 is on the next page.

Question 2.

Let $\bar{Y}$ denote the average number of children ever born to women aged 15-49 years in some population where a survey has been done. Three stage cluster sampling with stratification in the first stage was used to select the sample of women, 8,025 of whom participated in the survey. Properly computed estimates of $\bar{Y}$, its standard error, and its "design effect" are 2.84, $4{,}24 \times 10^{-2}$ and 1.33, respectively.

a) Ignoring the finite population correction, estimate the element variance (S^2) for the Y-variable from these results.

b) Calculate the coefficient of variation for the estimate of $\bar{Y}$.

c) Suppose that, in adding values of the Y-variable over all sample respondents and dividing by 8,025, you get 2.67. Assuming that no computational errors are involved, what does this tell you about the estimator used to compute 2.84 above and about the sampling design of this survey?

d) Suppose that simple random sampling had been used in this survey. What is your estimate of the sample size that would have been needed to achieve the same level of sampling error as reported above?

[illegible] and after going [illegible]

[illegible]	[illegible]	[illegible]	[illegible]	56	53	52	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	72	49	54	[illegible]	[illegible]

[illegible] there was significant change in weight. Let [illegible] time k where k = 1 before the diet, k = 2 afterward [illegible] β_1 = mean weight before the diet and β_2 = mean weight after [illegible]

[illegible] using a linear model:

$$Y = X\beta + \varepsilon$$

where [illegible] $Y_{2,1}, \ldots, Y_{2,10}$), $\beta' = (\beta_1, \beta_2)$,

[illegible] independent normal $(0, \sigma^2)$

[illegible] H_0 that there was no change in weight, that is, $\beta_1 = \beta_2$.

[illegible] H_0 is true then $E(D_j) = 0$, so another test [illegible]

$$D_j = \mu + \varepsilon_j$$

$$D_j = \varepsilon_j$$

[illegible] normal $(0, \sigma^2)$ for $j = 1, \ldots, 10$. Use this set up to test [illegible] $= 0$, that is, there was no change in weight.

[illegible] assumptions you made for the analyses you made in (i) and (ii). Which [illegible] more appropriate? Why?

Question 3.

A. Briefly describe or explain the following terms:
   (a) OS
   (b) TSO
   (c) Track on a magnetic disk
   (d) Block
   (e) Logical Record
   (f) Byte
   (g) JCL

B. Compare and contrast: (That is, demonstrate that you know what each of these terms means and what are the differences among them.)
   (a) OS dataset
   (b) SAS database (SAS data library)
   (c) SAS dataset

C. The printout in Figure 1 was produced by the PROC PRINT statement in line 00140002 of the SAS program shown in Figure 2. Show what will be printed as a result of

   (a) the PROC PRINT statement in line 00160002. (Write out the entire output to be produced by SAS except titles and page headings.)

   (b) the PROC MEANS in statement 00180002 and related statements.

   (c) the PROC PRINT statement in line 00380002.

   (d) the PROC PRINT statement in line 00470002.

Figure 1

OBS	NAME	SEX	AGE	HEIGHT	WEIGHT
1	ALFRED	M	14	69	112
2	ALICE	F	13	56	84
3	BERNADETTE	F	13	65	98
4	BARBARA	F	14	62	102
5	JAMES	M	12	57	83

Question 4.

Suppose that the regression function of Y on a real-valued x is continue; suppose also that it is linear on $x < 0$ and $x \geq 0$ separately, but with possibly different slopes on these two intervals. That is, we can write

$$Y = \mu(x; \beta) + \varepsilon$$

where

$$\mu(x; \beta) = \beta_1 + \beta_2 x, \quad x < 0,$$
$$= \beta_1 + \beta_3 x, \quad x \geq 0.$$

An experiment is conducted in which one measurement of Y is made at each of 21 settings x_i, $i = 1, \cdots, 21$, which start at -10 and increase by one unit at a time (so that $x_{21} = 10$). Assume $Y_1, \cdots, Y_{21}$ are independent and normally distributed with constant variance $\sigma^2 > 0$.

(a) Determine the design matrix X and X'X.

The inverse of the X'X matrix is

$$(X'X)^{-1} = \frac{1}{783475}\begin{pmatrix} 148225 & 21175 & -21175 \\ 21175 & 5060 & -3025 \\ -21175 & -3025 & 5060 \end{pmatrix} .$$

Use this explicit part in solving the reamining parts of the problem. Also use the exact distributions rather than normal approximation.

(b) Determine explictly, in terms of $\hat{\beta}_3$ and $\hat{\sigma}$, the two-sided 95% confidence interval for β_3.

(c) Determine explictly the the two-sided 95% confidence interval for $\beta_3 - \beta_2$.

(d) Describe explicitly in terms of $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\sigma}$ how to obtain the p-value for the two-sided test of the hypothesis that $\beta_3 - \beta_2 = 0$; or equivalently, thath the regression function is globally linear.

(e) Describe explicitly in terms of $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\sigma}$ how to obtain the p-value for the two-sided test of the hypothesis that $\beta_3 = \beta_2 = 0$; i.e., that the regression function is globally constant.

(f) Let $s^2 = \Sigma_1^{21}(Y_i - \bar{Y})^2/20$ denote the sample variance fo the Y_i's. Describe explicitly in terms of s and $\hat{\sigma}$ how to obtain the p-value for the two-sided test of the hypothesis stated in (f).