1967 - 1980

DEPARTMENT-WIDE DOCTORAL WRITTEN EXAMINATIONS

of the

DEPARTMENT OF BIOSTATISTICS
School of Public Health
University of North Carolina at Chapel Hill

VOLUME I
Closed-Book Parts

Dana Quade and Michael J. Symons

Institute of Statistics Mimeo Series #1343

June 1981

1967 - 1980

DEPARTMENT-WIDE DOCTORAL WRITTEN EXAMINATIONS

of the

DEPARTMENT OF BIOSTATISTICS
School of Public Health
University of North Carolina at Chapel Hill

VOLUME I

CLOSED-BOOK PARTS

assembled and edited by

DANA QUADE

and

MICHAEL J. SYMONS

June 1981

TABLE OF CONTENTS

* for one student undergoing a special re-examination

INTRODUCTION

This publication (Mimeo Series #1343) and its companion (Mimeo Series #1344) contain the general written examinations which the Department of Biostatistics has set for doctoral degrees: the closed-book parts here and the take-home parts in #1344.

The first Department-wide doctoral-level written examinations were given in September 1967, in connection with the new PhD program in Biostatistics. (Prior to that time the Department had offered the PhD only in Public Health, for which it had not required any Department-wide examination.) The original format was: (A) Basic Written Examination, with two Parts, (I) Theory and (II) Analysis, each including 5 closed-book questions, of which candidates were to complete 4 in 3 hours; and (B) Advanced Written Examination, again with two Parts, (I) General and (II) Specialty Areas, each one week take-home, where Part I had five questions and Part II ten, and candidates were to complete 4 from each Part.

In the 1969-1970 academic year it was decided to drop the Advanced Written Examination and expand the Basic Written Examination to three Parts: (I) Theory, with 5 closed-book questions of which 4 were to be completed in 3 hours; (II) Applications, with 5 take-home questions of which 4 were to be completed in one week; and (III) Specialty Areas, with 6 closed-book questions of which 3 were to be completed in 3 hours. The Basics were given twice in this format, in February/March and again in June of 1970.

On 3 March 1971, the Department voted to abolish Part III of the Basic Written Examinations, and to let the Literature Review for the Dissertation satisfy the Graduate School requirement for a Preliminary Written Examination. This action produced a format for the Basics which has continued essentially without change until the present time. In 1972 the time allowed for Part I (closed-book) was expanded from 3 hours to 4. Also, the following instruction was added to the rubrics for both Parts:

> Most questions should be answered in the equivalent of less than 7 typewritten pages (300 words per page) and under no circumstances will more than the first 15 typewritten pages or the equivalent be read by the grader.

This rubric was dropped from Part I in 1974. It continues for Part II, and indeed changed from 7 and 15 pages to 5 and 10 starting in 1978; but it has never been strictly enforced.

It may be noted that the Basics were originally intended for the PhD only, but since 1979 candidates for the DrPH have also been required to pass Part II (although not Part I.) Note further that the Basics have served not only as a screening examination for the PhD, but also as the Master's Written Examination for the MS; hence each candidate was evaluated using one of three grades: pass (at doctoral level), fail at doctoral level but pass at master's level, or fail. This last practice continued until 1980, when it was decided to combine the MS examination with the MPH/MSPH examination.

The Basic Examination is given annually; originally the regular date was in January or February, but starting in 1979 it is in August. Special examinations have been given on several occasions when there were sufficient students to warrant it; the current regulations require this only if there are 5 or more students to <u>retake</u> a Part.

The Examinations Committee prepares and conducts all Department-wide written examinations, and handles arrangements for their grading. 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 of the Basic Examination counting separately in this context). The members of each grading team are to 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 is to include comments on serious shortcomings in any candidate's answer.

On the basis of a candidate'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. The faculty has never specified in advance what the passing score would be, out of the possible 200 total points per Part, but has set cut-points varying from about 100 to 140, usually a little higher for Part II than for Part I. 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 Examination 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. When performance was not of the standard required, candidates are reexamined at a later date set by the Examinations Committee.

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 DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 A.M., September 16, 1967)

1. Let X be the number of successes in n independent Bernoulli trials with probability p of success on each trial. Under appropriate conditions, the distribution of X is approximately Poisson. State and prove the limit theorem pertaining to this.

2. (a) Derive the density function of the ratio of two independent standard normal random variables:

$$U = \frac{X}{Y}, \qquad X \text{ is } N(0,1),\ Y \text{ is } N(0,1);\ X \text{ and } Y \text{ independent.}$$

   (b) What is the name of this distribution?

   (c) What is $P(U < 0)$?

3. (a) State and prove the Neyman-Pearson Lemma.

   (b) To what extent is this Lemma applicable in significance tests involving a composite hypothesis?

   (c) Name and define a test criterion applicable to cases where the Lemma can not be used.

4. Statistician A decided to observe a certain random process for one hour; on doing so, he recorded that 50 events occurred during that hour. Independently, Statistician B decided to observe the process until 50 events occurred; on doing so, he recorded that one hour was required for that to happen.

(a) Find the maximum likelihood estimate of the mean rate of occurrence of events, using A's data; using B's data.

(b) Are these estimates unbiased? What are their standard errors? If they are not unbiased, show how to correct them for bias.

(c) An onlooker remarked that

(i) "A and B have really observed the same thing, i.e., 50 events in one hour" and

(ii) "When 2 statisticians have observed the same thing, they should make the same estimates."

Discuss these points carefully.

5. X_1 and X_2 are independent and identically distributed random variables with density function

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{if } x > \theta , \\ 0 & \text{otherwise; } -\infty < \theta < \infty . \end{cases}$$

Show that $\min(X_1, X_2)$ is a sufficient statistic for θ, and find its density function.

BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(1:30 P.M., September 16, 1967)

EDITORIAL NOTE: Appended to this examination were:

1) Density function and moment-generating function for uniform, normal, exponential, gamma, chi-squared, and student's t distributions;
2) Table of standard normal distribution function $\Phi(z)$ for $z = 0(.01)2(.02)3(.2)4(.5)5$ and $\Phi^{-1}(p)$ for $p = .00001(\text{various}).1(.05).9(\text{various}).99999$.

1. In order to determine whether the addition of an anticariogenic agent to a dentifrice makes an appreciable difference in flavor, it was decided to run a duo-trio test with ten panelists.

   The duo-trio test consists of presenting each judge with a known standard (the present dentifrice) and two unknowns (the new dentifrice and the present dentifrice). Each judge is then asked to state which of the unknown dentifrices differs from the standard.

   (a) What is the exact probability that 8 or more panelists will be right when all are actually guessing?

   (b) What is the exact probability that 9 or more panelists are right when all are actually guessing?

   (c) What is the probability that all 10 are right when all are actually guessing?

   (d) If we want $\alpha \leq .05$, how many panelists must be right to reject the null hypothesis that all panelists are guessing?

   What is the alternate hypothesis? Is this a two-tail test? Explain.

   (e) If it is desired to assure 90% probability of detecting a true 2:1 chance of correct identification, while keeping $\alpha \leq .05$, how many judges must be used?

2. A test has been conducted to compare the effectiveness of a new deodorant product (A) with the effectiveness of a competitive product (C). Effectiveness is measured in terms of underarm odor scores at a fixed time after use of the product. The objective is to reach one of the following conclusions:

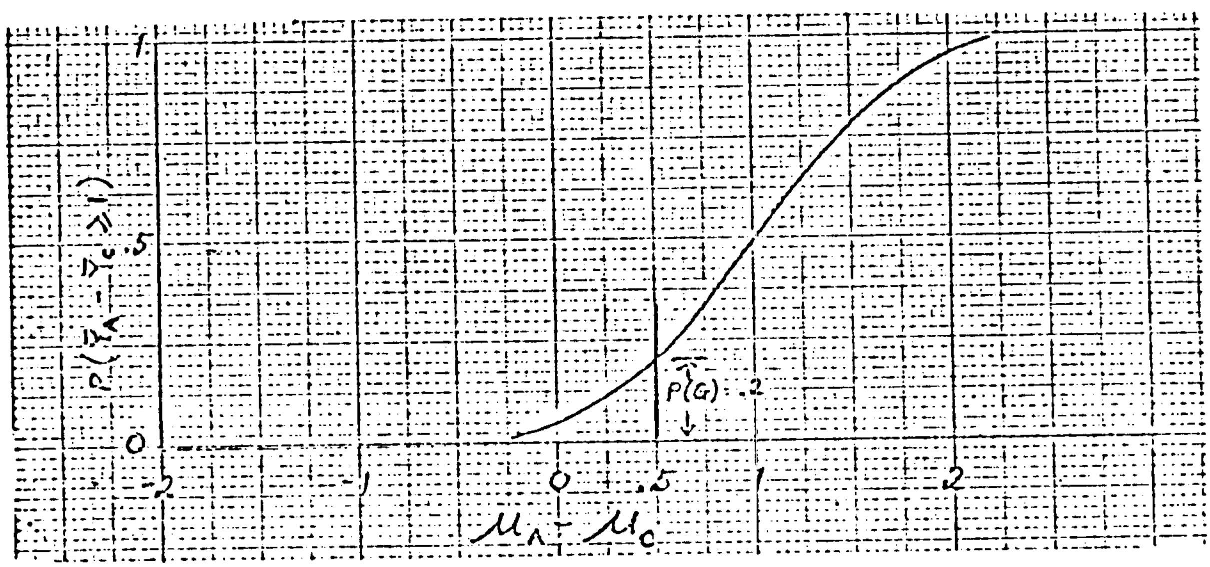
   (1) A is good enough to justify a formal claim that A is better than C.

   (2) A is not good enough to justify such a claim.

   Scores are defined so that they increase as effectiveness increases. The true mean scores for products A and C are μ_A and μ_C respectively. In the standard test μ_A is estimated by $\bar{Y}_A$ and μ_C is estimated by $\bar{Y}_C$. Conclusion (1) is reached if $(\bar{Y}_A - \bar{Y}_C)$ is at least 1.0 unit.

   In Figure 1 the performance of a single test is presented in graphical form. The plotted curve represents the probability of reaching conclusion (1) as a function of the true difference $\mu_A - \mu_C$.

Figure 1. Probability of Reaching Conclusion (1)



   (a) What is the null hypothesis in this test?

   (b) What is the α-risk? (Give numerical value.)

   (c) If a difference of 1.8 units is considered important, what is the β-risk? (Give numerical value.)

(continued)

For the remaining questions we assume that $\mu_A - \mu_C = .5$. Define two events as follows:

$$G \text{ is the event that } \bar{Y}_A - \bar{Y}_C \geq 1,$$

$$H \text{ is the event that } \bar{Y}_A - \bar{Y}_C \geq 0.$$

From Figure 1 the probability of occurrence of G is P(G) = .2. The probability of occurrence of H is known to be P(H) = .8.

(d) What is the conditional probability that we will reach conclusion (1) given that $\bar{Y}_A$ is at least as large as $\bar{Y}_C$?

(e) What is the expected value of the random variable $(\bar{Y}_A - \bar{Y}_C)$?

(f) Suppose two independent tests are run.

(i) What is the conditional probability that $\bar{Y}_A - \bar{Y}_C \geq 1$ in the second test given that we reached conclusion (1) on the basis of the first test?

(ii) What is the joint probability of reaching conclusion (1) on the basis of the first test and then finding $\bar{Y}_A - \bar{Y}_C < 1$ in the second test?

3. Suppose the following experiment was performed.

	Time Period			
Production Unit	1	2	3	4
1	AB	Ab	aB	ab
2	Ab	aB	ab	AB
3	aB	ab	AB	Ab
4	ab	AB	Ab	aB

Note that the treatments form a 2^2 factorial, where

a represents the low level of 1st treatment,

A represents the high level of 1st treatment,

b represents the low level of 2nd treatment,

B represents the high level of 2nd treatment.

(a) Exhibit the complete breakdown of degrees of freedom in the ANOVA for this experiment.

(b) Comment on the interactions which may be investigated in this experiment.

4. Given below are real data from a recent study to compare certain dental treatments. The children at a North Carolina orphanage were divided at random into 2 groups, and those in one group were given one treatment while those in the other group were given a different treatment. For each child a determination was made of the DMF rate (number of decayed, missing, or filled teeth) at the beginning of the experiment and two years later.

   (a) Make a comparison of the treatments, including a test of significance, using a method of analysis so simple that you can complete it <u>now</u>.

   (b) Describe briefly how you might attack a problem like this if you had plenty of time and computational assistance (and if there were more data, so that an elaborate analysis would be worthwhile).

Treatment 1					Treatment 2				
Child #	Sex	Initial Age	Initial DMF Rate	Final DMF Rate	Child #	Sex	Initial Age	Initial DMF Rate	Final DMF Rate
1	M	15	13	15	1	M	13	7	*
2	F	7	4	*	2	M	11	4	9
3	F	15	9	*	3	F	15	19	*
4	M	12	18	22	4	M	10	4	*
5	M	12	4	8	5	F	9	5	*
6	M	13	7	10	6	M	13	7	*
7	M	16	14	18	7	F	11	7	10
8	F	14	11	13	8	M	18	21	*
9	M	9	2	*	9	F	14	25	*
10	F	17	10	*	10	F	15	17	17
11	M	15	6	8	11	F	12	17	*
12	M	13	9	*	12	M	13	3	7
13	M	11	4	*	13	F	17	9	*
14	M	16	10	11	14	F	11	9	11
15	M	17	8	12	15	M	10	3	*
16	M	16	9	12	16	M	16	12	*
17	M	8	2	4	17	F	7	1	5
18	F	14	15	18	18	F	11	3	7
19	F	18	12	*	19	M	10	6	7
20	M	6	3	4	20	M	6	0	0
21	M	10	5	*					
22	M	16	8	10					

*Not available; child left orphanage before end of 2-year period.

5. It has been suggested that a linear relationship exists between the amount of mail handled in a post office and the man-hours required to handle it. In order to test out this hypothesis the following data were collected:

Four-week period, fiscal year 1962	Pieces of mail handled (X) (in millions)	Man-hours used (Y) (in thousands)
1	157	572
2	161	570
3	169	599
4	186	645
5	183	645
6	184	671
7	268	1053
8	184	655
9	180	637
10	188	667
11	184	656
12	182	640
13	179	609

Assume the following model:

$$Y = \beta_o + \beta_1 X + \epsilon \ ,$$

where X is a fixed (independent) variate, and ϵ is $N(0, \sigma^2)$

Calculations

$$\Sigma X = 2405 \qquad \Sigma X^2 = 453517$$
$$\Sigma Y = 8619 \qquad \Sigma Y^2 = 5892485$$
$$\Sigma XY = 1633249$$

$$\frac{\Sigma X}{N} = 185 \qquad \frac{\Sigma Y}{N} = 663$$

$$\Sigma X^2 - \frac{(\Sigma X)^2}{N} = 8592, \qquad \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} = 38734, \qquad \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 178088$$

(a) Calculate the ANOVA table for the assumed model using the above figures.

(b) Assuming all the above calculations are correct, test the hypothesis that $\beta_1 = 0$ using an α-risk of 5%.

(c) Comment on the validity of the error term in the Analysis.

(d) How do you explain that the best estimate of the intercept is negative?

(e) Is there any evidence here that the linear model is inadequate? Justify your answer.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12 A.M., September 21, 1968)

1. Define the term "sufficient statistic". Explain the importance of sufficient statistics in the theory of estimation: prove your statements as far as you can.

   Prove that the most powerful test of a simple null hypothesis against a simple alternative is based on a sufficient statistic, if such a statistic exists.

2. Find the probability that, in a random sample of size 3 from an exponential distribution, the mean will exceed the median.

3. Assume that a curve of the form $y = b_0 + b_1x + b_2x^2$ has been fitted and that the usual assumptions of normality hold. Derive a confidence interval for that x which produces the maximum or minimum value of y.

4. (a) Assume X and Y are random variables with means μ_x and μ_y $(\mu_y \neq 0)$, variances σ_X^2/n and σ_Y^2/n, and covariance σ_{XY}/n. Derive the asymptotic variance of X/Y.

   (b) Assume that $\underline{X}$ is a $(1\times p)$ random vector with mean vector $\underline{\mu}$ and variance matrix $\underset{p\times p}{\Sigma}$: define $q < p$ functions of the elements of X, viz. $f_i(\underline{X})$, $i = 1,2,\ldots,q$, which have derivatives up to second order. Derive the asymptotic covariance matrix of the vector $\underline{F} = [f_i(\underline{X})\ldots f_q(\underline{X})]$. For what kinds of functions is this expression exact?

5. Let X be a Poisson random variable with expected value θ.

   Let $Y_1, Y_2, \ldots,$ be independent variables each having the same expected value μ and variance σ^2.

   Let

$$Z = \begin{cases} \sum_{j=1}^{X} Y_j & X > 0 \\ 0 & X = 0 \end{cases}.$$

   Show that the correlation coefficient between X and Z does not depend on θ.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

(1:30-4:30 P.M., September 21, 1968)

1. Suppose a population consists of four strata, with the following weights: $W_1 = 10\%$, $W_2 = 50\%$, $W_3 = 25\%$, $W_4 = 15\%$. Suppose the variances within the strata are approximately: $s_1^2 = 16$, $s_2^2 = 4$, $s_3^2 = 9$, $s_4^2 = 12.25$. A sample of size n = 100 has been drawn from the population, with Neyman allocation used to determine the number of units selected from each stratum. What are these sample sizes? Suppose the observed means within the strata are $\bar{y}_1 = 10$, $\bar{y}_2 = 12$, $\bar{y}_3 = 14$, $\bar{y}_4 = 15$; and the observed within-strata variances are $s_1^2 = 15$, $s_2^2 = 5$, $s_3^2 = 10$, $s_4^2 = 12$. Give the appropriate estimate for the population mean; also give an estimate of the sampling variance of this estimate.

NOTE: Problem #2 is on the next page.

3. An investigator has approached you with the following data concerning the response of white rats to a certain drug. He tells you that he has examined 10 animals at each drug level, and the responses given are means. He has also calculated variances at each drug level.

$\bar{y}$ (mean response)	x (drug level)	s^2
5.2	1	1.40
3.9	2	.80
3.2	3	.65
8.4	4	3.30
7.5	5	3.00
6.8	6	2.70

Indicate what steps are needed to complete the analysis. How would you explain your results to this investigator?

2. An amateur weather forecaster made "rain" or "shine" predictions for 100 days. During this period he was successful on 8 out of his 20 "rain" forecasts and 63 out of his 80 "shine" forecasts.

This information might be organized for χ^2 computation in any of the following three ways. In each case the numbers in parentheses are the "expected frequencies" based on a null hypothesis which seems natural from the way the table is set up.

(a)

Successes	71	(50)	
Failures	29	(50)	$\chi^2 = 17.7$
Total	100		

(b)

	Prediction: Rain	Prediction: Shine		
Success	8(14.2)	63(56.8)	71	
Failure	12(5.8)	17(23.2)	29	$\chi^2 = 11.7$
			100	

(c)

		Predicted: Rain	Predicted: Shine		
Observed	Rain	8(5)	17(20)	25	
	Shine	12(15)	63(60)	75	$\chi^2 = 3.0$
		20	80	100	

For each case state the null hypothesis on which the expected frequencies are computed. Which of these three analyses leads to the most meaningful answer to the question: "Has the forecaster demonstrated real ability?" Justify your answer.

4. Two judges each ranked eleven contestants in a beauty contest, with the following results:

Contestant	Judge 1	Judge 2
A	6	4
B	4	1
C	6	7
D	1	5
E	2	2
F	6	6
G	10	9
H	3	3
I	9	11
J	11	8
K	8	10

Using three essentially different methods: obtain a measure of the agreement between the judges; interpret this measure; and test whether the judges agree significantly (i.e., calculate the test statistic and indicate exactly where and how you would look up the critical value).

5. A common practice in medicine is to use each patient as his own control when comparing two treatments A and B. That is: n/2 patients receive treatment A followed by B, and n/2 receive B followed by A. Assume that neither drug produces any appreciable residual effect and that the effect of treatment can be measured by a normally distributed random variable. Under what conditions is thus using each patient as his own control better, equal, or worse as a design than administering each of drugs A and B to a separate group of n patients?

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12 A. M., Saturday, February 21, 1970)

NOTE: Problem #1 is on the next page.

2. Suppose a certain population contains two strata of sizes $N_{1\cdot\cdot}$ and $N_{2\cdot\cdot}$ respectively. Suppose samples of sizes $n_{1\cdot\cdot}$ and $n_{2\cdot\cdot}$ are taken from the respective strata independently according to simple random sampling. Suppose that the selected elements are classified according to two binomial attributes, with one of them being the variable of principal interest such as occurrence of automobile accidents, and the other being a related covariable, such as age classified as under or over 35. The data appear as follows:

Stratum 1

Age	≤ 35	≥ 36	Total
Accident	n_{111}	n_{112}	$n_{11\cdot}$
No accidents	n_{121}	n_{122}	$n_{12\cdot}$
Total	$n_{1\cdot 1}$	$n_{1\cdot 2}$	$n_{1\cdot\cdot}$

Stratum 2

Age	≤ 35	≥ 36	Total
Accident	n_{211}	n_{212}	$n_{21\cdot}$
No accident	n_{221}	n_{222}	$n_{22\cdot}$
Total	$n_{2\cdot 1}$	$n_{2\cdot 2}$	$n_{2\cdot\cdot}$

a. Indicate what the appropriate estimates of the incidence rate of automobile accidents are for each stratum and the overall population when no information other than that given above is available. Also indicate an estimate of the standard error of these statistics.

b. How is the analysis in (a) changed if it is known that the population is subdivided as follows?

Stratum	1	2	Total
Age ≤ 35	$N_{1\cdot 1}$	$N_{1\cdot 2}$	$N_{1\cdot\cdot}$
Age > 35	$N_{2\cdot 1}$	$N_{2\cdot 2}$	$N_{2\cdot\cdot}$
Total	$N_{\cdot\cdot 1}$	$N_{\cdot\cdot 2}$	$N_{\cdot\cdot\cdot}$

1. A and B make independent measurements on a certain material having magnitude α. Both A's measurement and B's measurement are subject to random error, the former being uniformly distributed over $\alpha \pm 2$ and the latter uniformly distributed over $\alpha \pm 1$. What is the probability that A's and B's measurements will be within u of each other? What is the expected value of the magnitude of the discrepancy between the two measurements? (Hint: Take $U = |X-Y|$ where X is uniform on $[-2,2]$ and Y is uniform on $[-1,1]$.)

NOTE: Problem #2 is on the previous page.

3. Let F be a continuous distribution function with density function f and unique median η; let $X_1, X_2, \ldots, X_n$ be a random sample of n observations from F, and let Y_n be the sample median of the X's. (Assume n is odd.)

   a. Find the density function of Y_n, say g.

   b. Show that if f is symmetric about η then so is g.

   c. Show that if $E[X^k]$ exists for some $k > 0$ then so does $E[Y_n^k]$.

   d. Show that $\{Y_n,\ n=1,3,5,\ldots\}$ is a consistent estimator sequence for η.

4. Suppose we wish to estimate a set of Poisson probabilities. Using the maximum likelihood estimate of λ, we find the maximum likelihood estimates are

$$\hat{p}_i = \frac{e^{-\hat{\lambda}}\,\hat{\lambda}^i}{i!}, \quad i=0,1,2,\ldots \quad .$$

   a. Show that the mean of $\hat{p}_i$ is

$$\frac{e^{-n\lambda(1-e^{-\frac{1}{n}})}}{i!\,n^i}\, E[k^i]$$

   where k is a Poisson variable with parameter

$$\mu = n\lambda e^{-1/n} \; .$$

   b. Show that

$$\lim_{n\to\infty} E\,(\hat{p}_o) = e^{-\lambda} \; .$$

   c. What properties are associated with the estimates $\hat{p}_i$?

5. A traffic accident researcher found an average of 0.5 accidents per driver (for a certain time period). In his report he called attention to the fact that 40% of the accidents were "caused by" (i.e., were on the records of) only 9% of the drivers. Assuming an adequate sample was taken, would you accept this as evidence of "accident proneness" of a relatively small percentage of the drivers? Justify your answer.

Poisson distribution with $\lambda = 0.5$

x	p(x)
0	.6065
1	.3033
2	.0758
3	.0126
4	.0016
5	.0002
6	.0000
.	.
.	.
	1.0000

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART III

(9-12 A.M., Saturday, March 7, 1970)

Instructions

a) This is a closed-book examination.

b) The time limit is three hours.

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

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

e) Return the examination with a signed statement of the honor pledge.

1. A recently-reported Taiwan study used a "matching" study in an attempt to "estimate the real demographic impact of the IUD". The introduction and methods (design) sections of the paper are reproduced and attached*. The pre-IUD and post-IUD fertility data from the household register were used to obtain three year pre-IUD and post-IUD to December 31, 1967 fertility rates, which were compared to estimate the demographic impact.

   a) Discuss the advantages of this design as compared with a study of acceptors alone.

   b) Discuss the potential biases when using the two registers to identify IUD acceptors and non-acceptors in the household register system.

   c) Assume that acceptors and non-acceptors can be accurately and completely identified before "matching". Discuss how other factors including migration might bias the results of the present study.

   d) How might one analyze the available data to get some idea of whether biases exist?

* M.C. Chang, T.H. Liu and L.P. Chow, "Study by matching of the demographic impact of an IUD program", Milbank Memorial Fund Quarterly, April, 1969 (pp. 137-140 attached to examination).

2. In an experiment on the response of three variables to two drugs, the following data were obtained:

	Case	Variable A	Variable B	Variable C
Drug 1	1	5	10	8
	2	4	12	7
	3	6	12	9
	4	6	10	7
	5	4	11	9
Drug 2	6	10	12	7
	7	9	13	9
	8	9	12	9
	9	8	11	7
	10	9	12	8

$\underline{\bar{x}}_1 = (5,11,8)$

$\underline{\bar{x}}_2 = (9,12,8)$

$$S = 1/8 \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 3 \\ 0 & 3 & 8 \end{pmatrix} \qquad S^{-1} = 8 \begin{pmatrix} 1/6 & 0 & 0 \\ 0 & 8/39 & -1/13 \\ 0 & -1/13 & 2/13 \end{pmatrix}$$

a) Is the given inverse correct?

b) Assuming multivariate normality: which, if any, variables are independent? If multivariate normality is not assumed, what statements can be made about independence?

c) Test for equality of means in the two groups. (i.e., calculate the test statistic and indicate how to look up its significance).

d) Discuss the effect of sample size, and number of variables, on the test you have just performed.

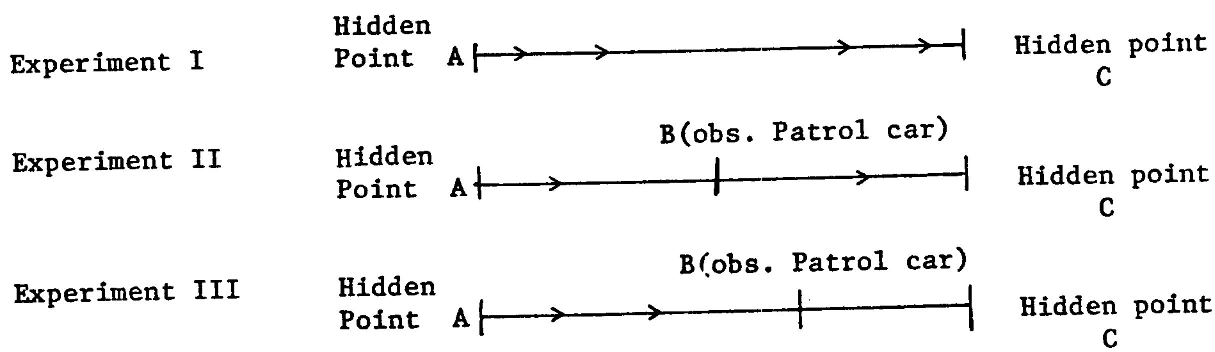
3. Assume that a health department serving a metropolitan area with a population of approximately a million persons has initiated a cervical cancer screening program aimed at reaching approximately 5000 women. You have been asked to serve as statistical consultant for the program, with specific responsibility for preparing a brief statistical report suggesting:

   a. Population groups the program might seek to reach, and the best location of program services for this purpose.

   b. Statistical data and analyses needed in program operation and evaluation.

Briefly highlight the type of report you might prepare. Include, for example, ways you might use mortality and any available morbidity data in determining target groups, types of statistics on program services that might be maintained, and means by which you would seek to measure program success.

4. The following experiments were undertaken to measure the effect of an observable stationary patrol car on the speed of vehicles. The method of measurement used consisted in placing hidden cars with Vascar speed measuring equipment at the extreme points of a section of highway to determine the actual speed of vehicles at these points. In the control experiment I, no observable patrol vehicle is used; in experiment II, an observable patrol car is midway between the two hidden points; and in experiment III, the observable patrol car is nearer the terminal hidden point. These may be diagrammed as follows:



where A is the initial point and C is the terminal point. The data were as follows:

Experiment		I			II			III		
Speed at C		≤60	>60	Total	≤60	>60	Total	≤60	>60	Total
Speed at A	≤60	19	8	27	139	15	154	46	4	50
	>60	1	4	5	89	38	127	19	8	27
Total		20	12	32	228	53	281	65	12	77

Discuss the analysis of these data in terms of a general categorical data model. Indicate the nature of the effects of the observable patrol car and differences among the experiments. Finally, give approximate values for test statistics directed at the various hypotheses of interest.

5. A test for the presence or absence of gout is based on the serum uric acid level x in the blood. Assume x is a normally distributed random variable. In healthy individuals x has a mean of 5.0 mg/100ml and a standard deviation of 1.0 mg/100ml and in diseased persons it has a mean of 8.5 mg/100 ml and a standard deviation of 1.0 mg/100ml.

A test T for the presence of gout is to classify those persons with a serum uric acid level of at least 7.0 mg/100ml as diseased.

(a) In epidemiologic terminology, the "sensitivity" of a test for a disease is defined as the probability of detecting the disease when it is present. Express the sensitivity of T as a test for gout in terms of standard tabulated functions.

(b) The "specificity" of a test is defined as the probability of the test not detecting the disease when it is absent. Express the specificity of T similarly.

Alternatively, this can be regarded as a statistical test of the null hypothesis

$$H_o: \ \mu = 5.0\text{mg}/100\text{ml}$$

against the alternative hypothesis

$$H_a: \ \mu = 8.5\text{mg}/100\text{ml}$$

The data is a single observation on the serum uric acid x of the person to be classified.

What is the rejection region of the test T?

What is the size of the test (expressed in terms of tabulated functions)?

Relate the epidemiological concepts of sensitivity and specificity to α, β, and $1-\beta$, the Type I and Type II error probabilities and the power of T respectively.

6. In his continuing research into ways of improving our environment, a scientist has developed a fly ointment which, when applied to the legs of young flies, may or may not inhibit (slow down) the growth of their legs. Since short-legged flies are known to have more trouble with takeoffs and landings (fall on their little faces...), if the ointment works the scientist feels it will be a boon to future generations of mankind.

In order to prove that his ointment does indeed inhibit growth of flies' legs the scientist performed the following experiment. Three concentrations of ointment were prepared, containing 0(control), 0.2 mg/ℓ and 0.4 mg./ℓ of the active ingredient. Thirteen young flies (6 females and 7 males) were captured and assigned to treatments at random. After being assigned to treatments, the flies in each group were numbered 1,2,...,n; as shown in the following table:

		Concentration		
Sex		0.0	0.2	0.4
M	cell # fly #	1 1,2,3,4	2 1	3 1,2
F	cell # fly #	4 1,2	5 1,2,3	6 1

For example, there were $n_3 = 2$ male flies who received the 0.4 concentration ointment.

Before the ointment was applied, the lengths of each fly's fore and hind legs were measured and recorded. The appropriate ointment was then applied to the right fore and hind legs of each fly, and the flies were allowed to mature under controlled conditions, after which their legs were re-measured.

The experimenters chose to analyze the data with the following model:

$$Y_{ij} = \mu_i + \alpha X_{ij} + e_{ij}, \quad j = 1,2,\ldots,n_i \quad i = 1,2,\ldots,6 \tag{1}$$

where:

Y_{ij} is the difference between <u>growth</u> of the right foreleg (which received the ointment) and the growth of the left foreleg of the j-th fly in the ith cell.

X_{ij} is the average length of the two forelegs of the jth fly in the ith cell before the ointment was applied.

continued

6. continued

Questions:

a. In the model (1) above, what is the interpretation of the μ_i's?

b. In the model above, what is the interpretation of the e_{ij}'s?

c. What assumptions about the e_{ij}'s are necessary if we are to use the Gauss-Markov theorem to justify using least squares estimators of α and the μ_i's?

d. What further assumptions about the e_{ij}'s are necessary if we want to perform the usual tests of hypothesis?

Let $\underline{\beta} = (\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\alpha)'$.

We can test hypotheses of the form $C\underline{\beta} = \underline{0}$. Write down the C-matrices for testing the usual hypotheses of an analysis of covariance. That is, write down the C-matrices to test:

e. For a "Sex" main effect;

f. For a "Concentration" main effect;

g. For a Sex x Concentration interaction.

h. Whether $\alpha = 0$.

i. Write a C-matrix to test whether there is a significant linear trend in response, y, as concentration increases.

j. Write a C-matrix to test whether there is a significant quadratic trend in the response, y, as concentration changes. Write this C-matrix so that the resulting test statistic is not correlated with the test statistic generated by the test in part i.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12 A.M., Friday, June 26, 1970)

Instructions

a) This is a closed-book examination.

b) The time limit is three hours.

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

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

e) Return the examination with a signed statement of the honor pledge.

1. In a certain population the number of children per family has a distribution with mean 3 and variance 6; and every child, independently of all others, has probability 1/2 of being male. Let a male child be chosen at random from this population, and let the random variable X be the number of brothers he has. Find $E[x]$.

2. For each of the following distributions: write down the probability function; specify a model, including all necessary assumptions, which would produce data having the distribution; and give an illustrative example from the biological or health sciences.

   a) Negative binomial

   b) Multinomial

   c) Hypergeometric

3. Suppose X_1 is exponential with mean λ_1 and X_2 is exponential with mean λ_2. What is the distribution of X_1/X_2? What is the mean of this distribution? Find an expression for the $p^{\underline{th}}$ percentile.

4. Given independent random samples, of size n each, from three normal populations having a common known variance σ^2 but unknown means:

$$\{X_{1i},\ i=1,2,\ldots,n\} \quad \text{from } N(\mu_1,\sigma^2),$$
$$\{X_{2i},\ i=1,2,\ldots,n\} \quad \text{from } N(\mu_2,\sigma^2),$$
$$\{X_{3i},\ i=1,2,\ldots,n\} \quad \text{from } N(\mu_3,\sigma^2),$$

construct a most powerful test of size α for testing the null hypothesis

$$H_0:\ \mu_1 = \mu_2 = \mu_3 = 0$$

against the alternative hypothesis

$$H_A:\ \mu_1 = a_1,\ \mu_2 = a_2,\ \mu_3 = a_3,$$

where a_1, a_2, a_3 are given constants. Be precise as to details of the critical region.

For the specific case $\sigma = 2$, $a_1 = .3$, $a_2 = .4$, $a_3 = 1.2$, $\alpha = 0.025$, find the smallest sample size required to ensure that the power of the test against H_1 will be not less than 0.975.

5. In a certain region a 15% random sample of hotels is taken and classified according to whether or not every bedroom has a telephone. The results are:

Total Number of bedrooms	Hotels in Region	Hotels in Sample	Sample hotels with full number of telephones
1-5	400	80	2
6-20	1500	200	50
21-50	600	100	45
51-100	300	50	35
101-	200	20	18
Total	3000	450	150

Compute the appropriate unbiased estimates of the total number of hotels with telephones in every room and the corresponding variance, assuming the data arose from (a) a simple random sample; (b) a stratified random sample as indicated in the table. Find the relative efficiency of the two methods of selection.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART III

(9-12 a.m., Friday, July 10, 1970)

INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is three hours.

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

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

e) Return the examination with a signed statement of the honor pledge.

1. Briefly specify

   a. Some ways in which vital statistics might be used in development and evaluation of programs for detection and care of chronic illness, such as cancer and heart disease -- including specific reference to some analytic techniques that might be employed.

   b. Several ways in which other types of health statistics, e.g. on morbidity and on health services provided, might be used in development and evaluation of the chronic disease programs.

   c. Some special needs for improved analytic techniques that might be used in applying vital or other health statistics in development and evaluation of chronic disease programs.

2. The following data were obtained from a sample of motorcycle operators on campus in order to investigate the dependence of average daily mileage on the operator's class status and marital status.

Mileage	Very Low	Low	Medium	High	Sample Size
Undergraduate Single	39	39	42	49	169
Graduate Single	3	4	9	6	22
Undergraduate Married	3	3	0	3	9
Graduate Married	15	13	9	2	39

Discuss the statistical analysis of these data in terms of a model which accounts for the relative importance of the marital status effects, class status effects, and marital status x class status interaction on mileage.

3. In this problem, $N_p(\underline{\mu},V)$ refers to the p- variate normal distribution with mean vector $\underline{\mu}$ (pxl) and variance-covariance matrix V.

(a) What is the joint frequency function (probability density function) of $N_p(\underline{\mu},V)$?

(b) What is the moment generating function of $N_p(\underline{\mu},V)$?

(c) If A is an mxp matrix ($m \leq p$) of rank m, and $\underline{y}$ has the $N_p(\underline{\mu},V)$ distribution, what is the distribution of $\underline{z} = A\underline{y}$? Prove your answer.

(d) Prove: If $\underline{y}$ has the $N_p(\underline{\mu},V)$ distribution then the components y_i are jointly independent if and only if cov $(y_i,y_j) = 0$ for all $i \neq j$, i.e., iff V is diagonal.

4. Consider n randomized blocks of $k(\geq 2)$ plots each where k different treatments are applied (once each). The response of the plot in the i^{th} block receiving the j^{th} treatment is $\underset{\sim}{X}_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})'$, where

$$\underset{\sim}{X}_{ij} = \underset{\sim}{\mu} + \underset{\sim}{\beta}_i + \underset{\sim}{\tau}_j + \underset{\sim}{e}_{ij};\ \underset{\sim}{e}_{ij} \sim N_2(0,\Sigma),$$

and Σ is positive definite.

(a) Show how to test the null hypothesis

$$H_o^{(1)} : \underset{\sim}{\tau}_1 = \dots = \underset{\sim}{\tau}_k \text{ vs } \underset{\sim}{\tau}_j \neq \underset{\sim}{\tau}_{j'} \text{, for at least one } (j,j').$$

(b) Treating $X^{(2)}$ as the concomitant variate and assuming that $\beta_i^{(2)} = 0,\ \forall_i$, and $\tau_j^{(2)} = 0,\ \forall_j$, show how to carry out the analysis of covariance test for the null hypothesis

$$H_o^{(2)} : \tau_1^{(1)} = \dots = \tau_k^{(1)} \text{ vs } \tau_j^{(1)} \neq \tau_{j'}^{(1)} \text{, for at least one } (j,j').$$

In the second problem, what is a simultaneous test for the paired differences $\tau_j - \tau_{j'}, 1 \leq j < j' \leq k$?

5. A "gross nuptiality table" is a "life table" in which a population of single individuals suffers decrementation through marriage in the absence of mortality.

a) Describe such a table including the assumptions involved in constructing it.

b) Derive an expression for the average age at marriage in the stationary life table population.

c) Show how mortality can be incorporated as an additional source of decrementation to arrive at what is called a "net nuptiality table".

6. The following data on the incidence of pneumococcal pneumonia have been collected in Chapel Hill during the past year.

Age at last birthday years	Boys			Girls		
	Cases	Pop.	Attack rate (per 1000)	Cases	Pop.	Attack rate (per 1000)
Under 1	15	200	75.0	13	195	66.7
1	25	225	111.1	26	305	85.2
2	17	210	81.0	16	215	74.4
3	15	202	74.3	12	205	58.5
4	9	195	46.2	7	180	38.9

Suppose you are interested in learning whether there is a sex difference in susceptibility or exposure to the disease among children under five years of age.

a. List at least six different tests that might be performed to test this hypothesis.

b. Describe in general terms the power of each test or reasons why one would want to use it, or not use it, in the present instance.

c. Perform at least two of these tests, and indicate what inferences you can draw about a sex differential from each test.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12 A.M., Saturday, February 20, 1971)

1. Consider a finite population of size N with population mean $\bar{X}$ and population variance S^2. A simple random sample of size 3 is drawn with replacement from this population. Let $\bar{x}'$ denote the unweighted mean over the distinct units in the sample.

(7 points) a) Show that the probabilities that the sample contains 1, 2, or 3 distinct units are

$$P_1 = 1/N^2, \quad P_2 = 3(N-1)/N^2, \quad P_3 = (N-1)(N-2)/N^2.$$

(9 points) b) Show that, for estimating $\bar{X}$, $\bar{x}'$ is conditionally unbiased (i.e., conditional upon a fixed number of distinct units). Show also that $\bar{x}'$ is unconditionally unbiased for $\bar{X}$.

(9 points) c) Find the unconditional variance of the estimator $\bar{x}'$.

2. A bivariate random variable $\underset{\sim}{x} = (x_1, x_2)$ has a bivariate lognormal distribution if $\underset{\sim}{y} = (y_1, y_2)$, where $y_1 = \log_e x_1$ and $y_2 = \log_e x_2$, has a bivariate normal distribution. That is

$$f(\underset{\sim}{y};\underset{\sim}{\mu},\Sigma) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\underset{\sim}{y}-\underset{\sim}{\mu})\Sigma^{-1}(\underset{\sim}{y}-\underset{\sim}{\mu})'],$$

where $\underset{\sim}{\mu} = (\mu_1, \mu_2)$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$

(10 points) a) Give explicitly the distribution of $\underset{\sim}{x} = (x_1, x_2)$.

(10 points) b) What is the distribution of $z = x_1/x_2$?

(5 points) c) What is $E(z)$?

3. In a two-organ system (e.g., kidneys, lungs) assume that the life-times of the two organs are independent exponentials, each with density function

$$f(x) = \lambda_o e^{-\lambda_o x}, \qquad \lambda_o > 0,\ x \geq 0.$$

Assume also that as soon as one organ fails, at time x_o, say, the surviving organ has density function

$$g(y) = \lambda_1 e^{-\lambda_1 (y-x_o)}, \quad y \geq x_o$$

with $\lambda_1 > 2\lambda_o$

(5 points) a) Show that at time t both organs are still functioning with probability

$$P_o(t) = e^{-2\lambda_o t}, \quad t \geq 0.$$

(12 points) b) Show that at time t the system is still functioning on one organ with probability

$$P_1(t) = \frac{2\lambda_o}{\lambda_1 - 2\lambda_o} (e^{-2\lambda_o t} - e^{-\lambda_1 t}), \quad t \geq 0.$$

(8 points) c) Show also that the density function of system life is

$$h(t) = \frac{2\lambda_o \lambda_1}{\lambda_1 - 2\lambda_o} (e^{-2\lambda_o t} - e^{-\lambda_1 t}), \quad t \geq 0.$$

4. The following table records 292 litters of mice classified according to the size of the litter and the number of females in the litter.

	NUMBER OF FEMALES				
SIZE OF LITTER	0	1	2	3	4
1	8	12			
2	23	44	13		
3	10	25	48	13	
4	5	30	34	22	5

(10 points) a) With the assumption that for any specified size of litter the number of females is binomially distributed, show that there is no significant difference at the 5% level among litter sizes with respect to the proportion of females.

(15 points) b) Test the significance of deviations from the assumption of the binomial distribution (the numbers have been chosen so that the arithmetic is easy), and then comment (no calculations are required) on how your procedure for this test would require modification if the first test had shown a variation in the proportion of females with litter size.

5.

(8 points) a) Define the following terms:

i) Unbiased estimator

ii) Consistent estimator

iii) Efficient estimator

iv) Sufficient estimator

(6 points) b) Give an example of an estimator for a parameter of a normal distribution which is unbiased but not consistent and prove your result.

(5 points) c) Give an example of an estimator for a parameter of a normal distribution which is consistent but not unbiased and prove your result.

(6 points) d) Show that there is no sufficient estimator for the location parameter θ of the Cauchy distribution

$$f(x) = \frac{1}{\pi[1+(x-\theta)^2]} .$$

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12 A.M., Friday, June 25, 1971)

1. It is desired to select a sample from a certain population in order to estimate the mean of a particular variate, within ±2% of its true value with probability at least 95%. It is known from previous experience with this population that this variate has approximately the following distribution:

Variate Value Class Interval	Approximate Percentage
$0 < X \leq 10$	10%
$10 < X \leq 20$	20%
$20 < X \leq 30$	40%
$30 < X \leq 40$	15%
$40 < X \leq 50$	10%
$50 < X \leq 60$	5%

Using the above information, what sample size will be necessary in order to satisfy the objectives of the survey (at minimum cost where total cost is presumed to be a linear function of the number of observations)?

2. A 2 × 5 factorial experiment is to be replicated four times using 40 individuals chosen from a population of 10,000 individuals. Use this situation to discuss

   a) a completely randomized design

   b) a randomized block design (What might be used for blocking factors? What is randomized? Describe the differences in the sampling procedures from a completely randomized design.)

   c) a components of variance vs. a fixed effects model.

Give the mathematical model in each case with the appropriate formulas for the partitioned sum of squares and give the corresponding degrees of freedom.

3. Let X_1, X_2, X_3 be independent Poisson variables with parameters λ_1, λ_2, λ_3, respectively. (a) Obtain the UMP Test for

$$H_0: \ \lambda=\lambda_1+\lambda_2+\lambda_3=\lambda_0 \quad \text{vs.} \quad \lambda=\lambda^*>\lambda_0$$

(b) Suggest a test for

$$H_0: \ \lambda_1=\lambda_2=\lambda_3 \quad \text{vs.} \quad \lambda_i\neq\lambda_j \ \text{ for at least one } (i,j): \ 1\leq i<j\leq 3.$$

4. Suppose $X_1,X_2,\ldots,X_n$ $(n\geq 2)$ are mutually independent and normally distributed random variables with zero mean and unit variance. Let

$$U = \frac{X_1}{(X_1^2+X_2^2+\ldots+X_n^2)^{\frac{1}{2}}} \quad \text{and} \quad V = X_1^2+X_2^2+\ldots+X_n^2.$$

Show that the random variables U and V are mutually independent and find the distribution of U.

5. a) Find the distribution function of the largest observation in a random sample of size n from a distribution with c.d.f. F(x).

   b) Given a sample of size n from the uniform distribution on $[0,\theta]$, obtain a confidence interval for θ of confidence coefficient $1-\alpha$ based on the maximum likelihood estimator of θ.

RE-EXAMINATION FOR [redacted]

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 A.M. - 5 P.M., Friday, June 25, 1971)

Instructions

a) This is a closed-book examination.

b) The time limit is eight hours.

c) Answer four of the five questions on Part I of the Basic Written Examination and the two attached questions.

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

e) Put your code letter not your name on each question.

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

Special Question for [redacted] (I)

Consider a model of the form

$$E(y) = \alpha_1 e^{\beta_1 x} + \alpha_2 e^{\beta_2 x}$$

(a) State an appropriate error model and derive the least squares equations for estimating the parameters α_1, β_1, α_2, β_2.

(b) What problems arise in the least squares estimation of the parameters α_1, α_2, β_1, β_2?

(c) Suggest an approximate method of solving these equations.

(d) Suppose it is known that $\beta_1=\beta_2=\beta$. Consider an alternative error model and find an explicit formula for an estimator of β. You may assume $\alpha_1+\alpha_2=\alpha$ is known in this case.

Special Question for [redacted] (II)

We have a random sample of size n from a distribution whose probability density function is assumed to be

$$f_\theta(x) = \tfrac{1}{2}(1 + \theta x), \quad -1 \le x \le 1,$$

$f_\theta(x) = 0$ otherwise, where the parameter $\theta(-1 \le \theta \le 1)$ is unknown.

a) Find that unbiased estimator, T_o, of θ which is a linear function of the sample mean $\bar{X}$.

b) Show that if the true value of θ is 0, then no unbiased estimator has a smaller variance than T_o.

c) Show that T_o minimizes the maximum (with respect to θ) of the variance of any unbiased estimator of θ.

d) Do you notice an undesirable property of T_o?

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-1 A.M., Saturday, January 29, 1972)

1. Given a 2×2 (two way) analysis of variance design,

(10 points) a) Reformulate as a one way design by showing the relationships between the parameters of the two models.

(8 points) b) Show how the main effects and interactions may be tested by way of orthogonal contrasts applied to the one way model.

(7 points) c) Compare the Scheffe and Tukey methods of multiple comparisons discussing the relative advantages and disadvantages of each.

2. The concept of a hazard function is used in reliability. The <u>hazard function</u> $h(t)$ is defined such that $h(t)dt$ is the probability that the failure of a system will occur in the time interval t to $(t+dt)$ given that no failure occurred prior to time t.

(8 points) a) If $f(t)$ and $F(t)$ are the density function and cumulative distribution function of the random variable T, the time to failure, show that

$$h(t) = f(t)/[1-F(t)].$$

(10 points) b) Find $f(t)$ if $h(t) = \alpha\beta t^{\alpha-1}$ $(\alpha>0,\ \beta>0,\ t>0)$.

(5 points) c) Find the mean and variance of $f(t)$ found in (b).

(2 points) d) Show that the exponential distribution is a special case of this distribution.

3. Let x_i $(i=1,2,\ldots,n+1)$ be a random sample of size $n+1$ from a normal $N(\mu,\sigma^2)$ population.

(5 points) a) Prove that

$$\left(\frac{\Sigma(x_i-\bar{x})^2}{b},\ \frac{\Sigma(x_i-\bar{x})^2}{a}\right)$$

is a confidence interval for σ^2 which has confidence coefficient $1-\alpha$ provided

$$\int_a^b f_n(y)dy = 1-\alpha \ \ldots \ (i)$$

where

$$f_n(y) = \frac{y^{\frac{1}{2}n-1}\ e^{-\frac{1}{2}y}}{2^{\frac{1}{2}n}\ \Gamma(\frac{1}{2}n)}, \qquad y>0.$$

(10 points) b) If the interval is to be the shortest, show that a and b must satisfy (i) and

$$f_{n+4}(a) = f_{n+4}(b).$$

(10 points) c) Show that the likelihood-ratio test, at significance level α, of H_o: $\sigma^2=\sigma_o^2$ against H: $\sigma^2\neq\sigma_o^2$ is to accept H_o if

$$\Sigma(x_i-\bar{x})^2/\sigma_o^2$$

lies in the interval (a,b), where a,b now satisfy (i) and

$$f_{n+3}(a) = f_{n+3}(b).$$

[You may assume that $Y = \Sigma(X_i-\bar{X})^2/\sigma^2$ has density function $f_n(y)$.]

4. Suppose two independent samples of sizes n_1 and n_2 are taken from $N(\mu_i,\sigma^2)$ $i=1,2$. Denote then by $X_1,\ldots,X_{n_1}$, $X_{n_1+1},\ldots,X_{n_1+n_2}$ where the first n_1 come from population 1 and the remainder from population 2. For each X_i, an indicator variable Z_i is associated, where $Z_i=1$ if X_i is in population 1 and $Z_i=0$ if X_i is in population 2. Let r be the sample correlation between the X's and the Z's. Find the relation between r and the usual unpaired two sample t-test.

(25 points)

5. Discuss the relative merits of using personal interviews, telephone interviews, and mailed questionnaires as methods of data collection for each of the following situations:

   a) A television executive wants to estimate the proportion of viewers in the country who are watching his network at a certain hour.

   b) A newspaper editor wants to survey the attitudes of the public toward the type of news coverage offered by his paper.

   c) A city commissioner is interested in determining how homeowners feel about a proposed zoning change.

   d) A county health department wants to estimate the proportion of dogs that have had rabies shots within the last year.

(25 points)

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-1 A.M., Saturday, January 27, 1973)

1. In the analysis of family records on a certain inherited disease, the distribution of affected children is often given in the form

$$\Pr\{X=r\} = \binom{s}{r} \frac{\theta^r (1-\theta)^{s-r}[1-(1-\pi)^r]}{1-(1-\pi\theta)^s} \qquad (1)$$

for $r=1,2,\ldots,s$, where

s = the number of children in a family (family size);

θ = segregation probability;

π = ascertainment probability.

(10 points) (a) Show that (1) is a probability mass function.

(10 points) (b) Evaluate the expected number of affected children in a family of size s, using distribution defined in (1).

(5 points) (c) For small z, $(1-z)^k \doteq (1-kz)$. Using this fact, derive an approximation to (1) for small π. Show that the approximation is a probability mass function, and find its expectation.

2. Let X be a random variable with a gamma distribution, having density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \qquad 0<x<\infty, \quad \alpha,\beta>0.$$

Let $x_1,\ldots,x_n$ be a random sample from f.

(5 points) (a) If β is known, find a sufficient statistic for α.

(5 points) (b) If β is known, what is the best linear unbiased (BLUE) estimator for α? What is the variance of this estimator?

(5 points) (c) If α is known, find a sufficient statistic for β.

(5 points) (d) If α is known, β unknown, find the Cramer-Rao lower bound on the variance of an unbiased estimator of β. Find an estimator which achieves this lower bound. You may assume that a family of gamma laws with fixed α and unknown β is complete.

(5 points) (e) Find the estimator for the variance of f (α known, β unknown) which is optimal in the sense of part d).

3. Given a random sample $x_1,\ldots,x_n$ from a normal distribution $N(\mu,\sigma^2)$:

(10 points) (a) Construct the likelihood ratio test criterion for the hypothesis

$$H_o: \mu=\sigma^2$$

against alternatives

$$H_1: \mu\neq\sigma^2.$$

It is not necessary to reduce the test criterion to lowest terms.

(5 points) (b) Transform the test criterion to one with known asymptotic distribution, and show exactly how you construct a critical region for a test of level α.

(10 points) (c) Find the asymptotic variance of the maximum likelihood estimator of μ under H_o.

4. The Duchy of Grand Fenwick experiences a period of unprecedented economic growth from year 1 to year n_1 resulting from a massive American aid program. The end of such aid following year n_1 combined with a new policy of open immigration extended to citizens of underdeveloped countries, initiates a decreasing trend in per capita income from years n_1 to n, in stark contrast to the increasing trend experienced during the period of American aid. It is suggested that the change in per capita income, years 1 to n, may be appropriately described by two straight lines, necessarily intersecting at year n_1.

(5 points) (a) Let y_i = per capita income in year π.

$\underset{\sim}{y}$ = vector of y_i's, i=1,...,n

Define a parameter vector β and an appropriate X matrix for the model $E\underset{\sim}{y} = X\underset{\sim}{\beta}$ described above.

(5 points) (b) Give algebraic expressions for the elements of (X' X).

(3 points) (c) Give a matrix expression for $\underset{\sim}{\hat{\beta}}$, the least-squares estimates of $\underset{\sim}{\beta}$, (in terms of X and $\underset{\sim}{y}$).

(4 points) (d) Find $\underset{\sim}{\hat{\beta}}$ for the following data, where n_1=3

year	Per capita income
1	10
2	25
3	30
4	25
5	10

(4 points) (e) Suggest two things to look at in evaluating the fit of such a model.

(4 points) (f) Discuss briefly the effect of sampling considerations (how per capita income was measured) and possible distributional assumptions on the interpretation and use of this model.

5.

(8 points) (a) Show that for a non-negative random variable X for which $E|X|^k$ exists for some $k>0$,

$$EX^k = k\int_0^\infty x^{k-1}[1-F(x)]dx, \text{ where } F(x) = P\{X\leq x\},\ 0\leq x<\infty.$$

Hence, or otherwise, show that

$$EX = \int_0^\infty [1-F(x)]dx.$$

(7 points) (b) Let X be a non-negative integer valued random variable with $p_i = P\{X=i\}$, $i=0,1,2,\ldots,\infty$. Also, let

$$Q_i = P\{X>i\} = p_{i+1} + p_{i+2} + \ldots,\ i\geq 0;$$

$$p^{[q]} = p(p-1)\ldots(p-q+1).$$

Let then

$$\mu'_{[k]} = E(X^{[k]}) = \sum_{i=0}^\infty i^{[k]}p_i,\ k \text{ an integer } (\geq 1).$$

Show that

$$\mu'_{[k]} = k\sum_{i=0}^\infty i^{[k-1]}Q_i.$$

(10 points) (c) Let X have the Poisson distribution viz.,

$$P\{X=i\} = e^{-m}\, m^i/i!,\ i=0,1,\ldots,\infty;\quad m>0,$$

and let

$$\mu_r = E(X-m)^r, \text{ for } r=0,1,2\ldots\ .$$

Show that

$$\mu_{r+1} = mr\mu_{r-1} + m\frac{d}{dm}(\mu_r),\quad r\geq 1.$$

Hence, or, otherwise, evaluate the first four central moments of X.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9-12, Friday, May 25) 1973

1. Simple random samples drawn from each of three strata give data as follows:

Stratum I	Stratum II	Stratum III
10	21	30
12	23	31
13	25	28
17	23	40
8	20	38
9	26	37
20		37
11		33
13		
12		

(6 points) a) Estimate the population mean, and the variance of your estimate, under the assumption that the sample shown was drawn under proportional allocation.

(6 points) b) Estimate the population mean and the variance of your estimate, if the true stratum weights are equal.

(6 points) c) Assume the combined sample above was drawn by simple random sampling without stratification. Estimate the population mean and its variance.

(7 points) d) Now assume, under the conditions of (a), that data is available on a concomitant variable x, such that for each of the y values observed above, the corresponding x value is (y-5). Suppose also that the actual totals for the x variable in the population are known to be 8, 12, and 20 million in stratum I, II, III, respectively. Calculate the combined and seperate ratio estimates of the population total Y. Which would you prefer in this case, and why?

2. Let the probability p_n that a family has exactly n children be αp^n when $n \geq 1$, and $p_0 = 1 - \alpha p(1+p+p^2+\ldots)$. Suppose that the probability of giving birth to a boy is $\frac{1}{2}$ and that all births are independent events.

(15 points) a) Show that for $k \geq 1$ the probability that a family has exactly k boys is $2\alpha p^k/(2-p)^{k+1}$.

(10 points) b) Given that a family includes at least one boy, what is the probability that there are two or more?

NOTE: PRoblem #3 is on the next page.

4. Let $X_1, \ldots, X_n$ be n independent random variables distributed according to the $(0,\theta)$, $\theta>0$, rectangular distribution, and let $U_n = \max\limits_{1 \leq i \leq n} X_i$ and $V_n = \min\limits_{1 \leq i \leq n} X_i$. Show that

(7 points) a) U_n and V_n are stochastically independent,

(7 points) b) U_n is the maximum likelihood estimator of θ and is also a sufficient statistic for θ.

(3 points) c) Is U_n an unbiased estimator of θ?

(3 points) d) Obtain a 95% confidence interval for θ based on U_n.

3. Suppose that a random variable Y is related to a non-stochastic factor X over the interval $-1 \leq X \leq 1$ by the relationship

$$\eta = E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Let N pairs of observations (X_i, Y_i), $i=1,2,\ldots,N$, be available for which Var $(Y_i)=\sigma^2$, Cov $(Y_i, Y_{i'})=0$ if $i \neq i'$; further, let $\mu_k = \frac{1}{N} \sum_{i=1}^{N} X_i^k$, $k=1,2,3$, and assume that $\mu_1 = \mu_3 = 0$. It is desired to approximate η over $[-1,1]$ using

$$\hat{Y} = b_0 + b_1 X,$$

where $b_0 = \frac{1}{N} \sum_{i=1}^{N} Y_i$ and $b_1 = \sum_{i=1}^{N} X_i Y_i / \sum_{i=1}^{N} X_i^2$.

(5 points) a) Find $E(b_0)$ and $E(b_1)$.

(5 points) b) Show that, in general,

$$J = \int_{-1}^{1} E[(\hat{Y} - \eta)^2] dX$$

can be written as J=V+B where

$$V = \int_{-1}^{1} \mathrm{Var}(\hat{Y}) dX$$

and

$$B = \int_{-1}^{1} [E(\hat{Y}) - \eta]^2 dX.$$

(15 points) c) For the particular situation described earlier, show that V is minimized when $\mu_2 = 1$, that B is minimized when $\mu_2 = 1/3$, and that the value of μ_2 which minimizes J is the solution of a cubic equation.

NOTE: Problem #4 is on the preceding page.

5. Assume that an individual is subject to two types of failure and the two failure times are represented by random variables X and Y. The joint distribution of X and Y can be described in terms of the probability function

$$P[X>s \text{ and } Y>t] = e^{-\lambda_1 s-\lambda_2 t-\lambda_{12}\max(s,t)} \quad \text{for } s\geq 0,\ t\geq 0$$

$$= 1 \text{ for } s\leq 0,\ t\leq 0 \quad ,$$

where

$$\max(s,t) = s \quad \text{if } s\geq t$$

$$= t \quad \text{if } s\leq t$$

and $\lambda_1\geq 0$, $\lambda_2\geq 0$, $\lambda_{12}\geq 0$.

(8 points) a) Find the marginal distribution functions of X and Y.

(10 points) b) Find $P[X\leq x, \text{ and } Y\leq y]$

(7 points) c) Prove $P[X>x+u, \text{ and } Y>y+u \mid X>x \text{ and } Y>y]$

$$= P[X>u \text{ and } Y>u]$$

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 a.m.-1 p.m., Saturday, January 26, 1974)

1. Consider a population composed of the following

$$y_1=1,\ y_2=2,\ y_3=3,\ldots,y_{29}=29,\ y_{30}=30$$

(6 points) a. What is the variance of the mean for a random sample of size n=6 selected without replacement?

(6 points) b. What is the variance of the mean for the systematic sample which selects every 5th sampling unit from a random start?

(6 points) c. If the population were divided into 6 strata as follows:

$$(y_1,y_2,\ldots,y_5);(y_6,y_7,\ldots,y_{10});\ldots;(y_{26},y_{27},\ldots,y_{30}),$$

what is the variance of the estimate of the population mean for a stratified random sample of one observation per stratum?

(7 points) d. What is the efficiency of the three sampling schemes relative to simple random sampling?

2. The probability distribution of a random variable Y is given by

$$P[Y=y] = p\,q^{y-1}, \quad y=1,2,\ldots,$$
$$\text{and } q = 1-p.$$

(8 points) a. Show that the moment generating function of Y is given by

$$M(s) = (\lambda e^{-s} - \lambda+1)^{-1}, \text{ where } \lambda = \frac{1}{p}$$

Let Y_1 and Y_2 be two independent observations from the above distribution, and let $Z_1 = \max(Y_1,Y_2)$ and $Z_2 = \min(Y_1,Y_2)$.

(8 points) b. Find the probability distribution of Z_1.

(9 points) c. Find the conditional distribution of Z_1 given Z_2.

3. Let

$$Y_i = \alpha + \beta t_i + \varepsilon_i, \quad i=1,2,\ldots,n,$$

where $\varepsilon_i \frown N(0, \sigma_i^2)$, t_i is a known positive constant for every i, and α and β are parameters to be estimated. Obtain the maximum likelihood estimate of β when

(6 points) a. α is unknown and $\sigma_i^2 = \sigma^2$, $i=1,2,\ldots,n$.

(6 points) b. $\alpha=0$ and $\sigma_i^2 = \sigma^2$, $i=1,2,\ldots,n$.

(6 points) c. $\alpha=0$ and $\sigma_i^2 = t_i\sigma^2$, $i=1,2,\ldots,n$.

(7 points) d. $\alpha=0$ and $\sigma_i^2 = t_i^2\sigma^2$, $i=1,2,\ldots,n$.

4. During a holiday season, a merchant will make a net profit of α dollars for each unit amount of some commodity sold and will lose β dollars for each unit amount left unsold at the end of the season. Suppose that the merchant can purchase this commodity only before the season starts, and that the total amount his customers will want to buy during the season can be viewed as a continuous random variable with a given probability density function $f(x)$ for $x > 0$.

(17 points) a. If k is the number of units which the merchant should purchase in order to maximize the expected value of his profit G on the commodity, show that k is given by

$$F(k) = \frac{\alpha}{\alpha+\beta}\,.$$

(8 points) b. Compute k when $\alpha=1$, $\beta=2$, and

$$f(x) = .0002xe^{-.0001x^2}, \quad x>0.$$

5. Clinical studies in which several clinics participate in evaluation of therapies using a standard protocol have become relatively common. Assume an experimental design in which patients who meet protocol requirements are randomly assigned to treatments within clinics, and that the clinics that participate in the study are a random sample from the population of clinics that might use the therapies. An appropriate linear model might be

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where $E(y_{ijk}) = \mu+\alpha_i$; μ is the general mean, α_i is the differential effect of the i-th treatment, β_j is a random clinic effect, $(\alpha\beta)_{ij}$ is a random effect due to interaction between the j-th clinic and the i-th treatment, ε_{ijk} is a random effect due to the k-th patient in the i-th treatment and j-th clinic; β_j, $(\alpha\beta)_{ij}$, and ε_{ijk} are independently distributed, each with mean 0 and respective variances σ^2_β, $\sigma^2_{\alpha\beta}$, σ^2_ε for $i=1,\ldots,t$, $j=1,\ldots,c$; $k=1,2,\ldots,n_{ij}$.

(4 points) a. What is the variance of y_{ijk}?

(4 points) b. What is the covariance between two observations in the same clinic and treatment?

(4 points) c. What is the covariance between two observations in the same clinic, but different treatments?

(7 points) d. What is the covariance matrix for a vector of treatment means from a specified clinic?

(6 points) e. Find the covariance between $\sum_j c_{1j}\bar{y}_{ij}$ and $\sum_j c_{2j}\bar{y}_{ij}$, where $\sum_j c_{1j}=0$, $\sum_j c_{2j}=0$, $\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$, and i is fixed.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 a.m.-1 p.m., Saturday, May 18, 1974)

1. Suppose that X and Y are random variables with joint density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right\},$$
$$-\infty < x < \infty, \; -\infty < y < \infty.$$

6 points a) Obtain the moment generating function of the joint distribution of X and Y.

7 points b) Show that the correlation between X^2 and Y^2 is given by ρ^2.

6 points c) Find the distribution of

$$Q = (X^2-2\rho XY+Y^2)/(1-\rho^2)$$

6 points d) Show that $U = X+Y$ and $V = X-Y$ are independently distributed.

2. Let $Y_{(1)} < \ldots < Y_{(n)}$ be the ordered random variables of a sample of size n from the rectangular $(0,\theta)$ distribution, where $0 < \theta < \infty$. By a careless mistake, the observations $Y_{(k+1)},\ldots,Y_{(n)}$ were recorded incorrectly, so they were discarded. (Here, k is a positive integer less than n.)

10 points a) Show that the conditional distribution of $Y_{(1)},\ldots,Y_{(k-1)}$, given $Y_{(k)} = y$, is independent of θ.

8 points b) Hence, or otherwise, obtain the maximum likelihood estimator of θ and show that it is a function of $Y_{(k)}$ alone.

7 points c) If $k/n \to p$ as $n\to\infty$, for some $0 < p < 1$, what can you say about the asymptotic distribution of the maximum likelihood estimator of θ?

3. A bus line has length L. The probability density for the point X at which a passenger gets on the bus is proportional to $X(L-X)^2$, and the probability density for the point Y at which a passenger who entered at point X gets off the bus is proportional to $(Y-X)$. Find the probability that

5 points a) A passenger will get on the bus before point z,

5 points b) A passenger who got on the bus at point x will get off after point z,

15 points c) A passenger who gets off at point y got on before point z.

4. Suppose that N_1 represents the number of traffic accidents occurring at a certain highway location during the year prior to the installation of a certain improvement at that location, and that N_2 represents the number occurring during the year following the improvement. Suppose, moreover, that N_1 and N_2 are assumed to be independent Poisson random variables with means μ_1 and μ_2, respectively.

5 points a) What is the probability that $n_1 + n_2$ accidents occur during the two-year period?

6 points b) What is the conditional probability that n_2 accidents occur during the second year given that $n_1 + n_2$ occur during the two-year period?

7 points c) Show that the conditional probability of part (b) is the binomial probability $B(n_1; n_1 + n_2, \theta)$, where $\theta = \frac{1}{1+\rho}$ and $\rho = \frac{\mu_2}{\mu_1}$.

7 points d) Suggest a test of H_0: $\rho=1$ versus H_1: $\rho<1$ with approximate significance level α, and comment on any properties of your suggested test which you think are particularly good or bad.

5. Consider the following set of data:

	X	Y
GROUP A	3	12
	2	10
	1	10
	2	13
GROUP B	4	14
	3	14
	3	12
GROUP C	5	15
	1	8
	2	7
	3	10
	1	9

6 points a) Perform a simple linear regression of Y on X for the twelve bivariate observations given.

4 points b) For each observation, calculate the residual from the fitted regression line.

6 points c) Perform a one-way analysis of variance on the residuals, to see if the mean residuals differ significantly from group to group.

9 points d) Place the ANOVA test you have just performed in the context of general multiple regression analysis. (It is not necessary to get very technical.) Do parts a)-c) relate to a classical procedure familiar to you? Give a crucial restriction on the use of this technique.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 a.m.-1 p.m., Saturday, January 25, 1975)

1. Let $\psi(x)$ be a monotonic increasing function of x such that $\psi(x)\to 0$ as $x\downarrow x_o$ and $\psi(x)\to\infty$ as $x\uparrow\infty$. Let

$$F_X(x)=1-\exp[-\psi(x)],\ x>x_o,$$

be the cumulative distribution function (CDF) of a random variable X.

6 pts. a) Prove that $Y=2\psi(X)$ has a χ^2 distribution with 2 degrees of freedom.

3 pts. b) Let $X_1,X_2,\ldots,X_k$ be k (k>2) independent random variables, X_j having the CDF

$$F_{X_j}(x_j)=1-\exp[-\lambda u_j(x_j)],\ \lambda>0, j=1,2,\ldots,k,$$

where $u_j(x_j)$ satisfies the conditions of $\psi(x)$ in part (a) and $\lambda(>0)$ is a constant.

Using part (a), show that

$$Y = 2\lambda \sum_{j=1}^{k} u_j(X_j)$$

has a χ^2 distribution with 2k degrees of freedom.

10 pts. c) Suppose that λ is an unknown parameter and is estimated from the formula

$$\hat{\lambda} = (k-1)/\sum_{j=1}^{k} u_j(x_j).$$

Show that $\hat{\lambda}$ is an unbiased estimator of λ.

<u>HINT</u>: Note that $\hat{\lambda} = 2\lambda(k-1)/\chi^2_{2k}$.

6 pts. d) Find the variance of $\hat{\lambda}$.

2. Consider a simplistic model of morbidity (illness) for some ailment, say migraine headache. For any given calendar week, an individual belongs to one of two states:

Well (W) if he experiences no symptoms or signs of migraine headache at any time during the week;

Morbid (M) if he experiences symptoms or signs of migraine headache at some time during the week.

For a population P under study, assume that the probability that a randomly chosen individual will be well in week (k+1) depends only on his health statuses (W or M) in weeks k and (k-1). Denote the four probabilities for the status "well in week (k+1)" as below and assume that all probabilities are strictly between 0 and 1.

HEALTH STATUS WEEK (k-1)	WEEK k	PROBABILITY OF BEING WELL IN WEEK (k+1)
W	W	p_{ww}
W	M	p_{wm}
M	W	p_{mw}
M	M	p_{mm}

5 pts. a) Let the states of a Markov chain be defined by pairs $\underset{\sim}{z}=(z_{k-1}, z_k)$, where z_j is the health status of an individual in week j. Rename the states as: 1=(W,W); 2=(W,M); 3=(M,W); 4=(M,M). Let X_n be the state random variable. Verify that $\{X_n: n\geq 0\}$ is a homogeneous Markov chain over the state space {1,2,3,4}.

6 pts. b) Write out the one- and two-step transition matrices.

6 pts. c) Write down the equilibrium equations for this chain.

8 pts. d) Suppose that $p_{ww} = 0.8$, $p_{wm} = 0.2$, $p_{mw} = 0.4$, and $p_{mm} = 0.6$. Assume no emigration, immigration, births or deaths relative to the population P. After a large number of weeks, what is the probability that an individual selected randomly from P will be well in the calendar week of his selection? In the week following?

3. Consider a sequence of independent random variables $X_1, X_2, X_3, \ldots,$ each having the same expected value μ and standard deviation σ. The distribution of

$$Y_i = (X_i - \mu)/\sigma$$

does not depend on μ or σ, and

$$E\left[(Y_i - Y_j)^{-2}\right] = \lambda \ (<\infty) \text{ for } i \neq j.$$

Firstly, X_1 and X_2 are observed. Then, a further set of N observations $X_3, X_4, \ldots, X_{N+2}$ is taken, where N is the least integer not less than $M(X_1 - X_2)^2$, M being a fixed positive number. The quantity

$$\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_{k+2}$$

is then calculated.

7 pts. a) Show that

$$E(N) \geq 2M\sigma^2.$$

18 pts. b) Show that

$$\Pr\left[\,|\,\bar{X} - \mu| < \varepsilon\right] \geq 1 - \frac{\lambda}{M\varepsilon^2}$$

for any positive number ε.

4. It is required to take samples of sizes n_1 and n_2 from two normal populations having known variances σ_1^2 and σ_2^2, the purpose being to test the null hypothesis of equality of the population means μ_1 and μ_2 against the alternative $\mu_1 > \mu_2$.

8 pts. a) Setting

$$V = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

and using the notation which defines z_k by $\Pr(Z > z_k) = k$, find the expression for V which satisfies the conditions that the level of significance be α and the power of the test against the alternative $(\mu_1 - \mu_2) = \delta$ be $(1-\beta)$.

17 pts. b) For the V given in any such situation, find formulas for n_1 and n_2 (in terms of V, σ_1, σ_2) which will minimize sampling cost if the cost of making an observation in Population 1 is four times that in Population 2. What are the specific sample size values if $\sigma_1 = 5$, $\sigma_2 = 4$, $\alpha = .05$, and the power is to be .90 when $\delta = 3$?

5. Data is available of the form (X_i, Y_i, Z_i), $i=1,2,\ldots,n$.

5 pts. a) Consider the statistical model

$$Y_i=\alpha+\beta X_i+\varepsilon_i,$$

where α and β are fixed (unknown) parameters and the ε_i are independent random variables with zero mean and common variance σ^2. Give formulae for $\hat{\alpha}$ and $\hat{\beta}$, the least squares estimators of α and β. Give formulae for the variances of $\hat{\alpha}$ and $\hat{\beta}$. Derivations are not necessary.

12 pts. b) Consider the statistical model

$$Y_i= \alpha+\beta X_i+\gamma Z_i+\varepsilon_i,$$

where α,β and γ are fixed (unknown) parameters and the ε_i are independent random variables with zero mean and common variance σ^2. An analysis of variance for the data, under this model, is given below:

SOURCE	D.F.	S.S.	M.S.
X	1	8.0	8.0
Z/X	1	5.0	5.0
Residual	23	65.0	2.8
Total	25	78.0	

The correlation (Pearson) between Y and Z is $r_{YZ}= .4$. Calculate:

i) $r_{XY.Z}$, the partial correlation between Y and X, adjusting for Z;

ii) $r_{Y.XZ}$, the multiple correlation of Y with X and Z;

iii) r_{XZ}, the Pearson correlation of X with Z.

8 pts. c) Consider the statistical model

$$Y_i=\alpha^* + (\beta^*+Z_i^2)\, X_i+\varepsilon_i^*,$$

where α^* and β^* are fixed (unknown) parameters and the ε_i^* are independent random variables with zero mean and common variance σ^2. Derive formulae for $\hat{\alpha}^*$ and $\hat{\beta}^*$, the least squares estimators of α^* and β^*. What are the variances of these estimators?

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9a.m. - 1p.m., Saturday, January 24, 1976)

1. Let $\{X_i\}$ be a sequence of independent and identically distributed random variables and let $S_0 = 0$, $S_k = X_1 + \ldots + X_k$, $k \geq 1$. Also, let N be a non-negative integer-valued random variable, independent of the X_i.

   (a) Show that whenever the moments exist,

   $$E(S_N) = E(N)E(X_1)$$
   $$V(S_N) = E(N)V(X_1) + V(N)[E(X_1)]^2$$

   (b) Suppose that

   $$X_i = \begin{cases} 0, \text{ with probability } 1-p \\ 1, \text{ with probability } p \end{cases}, \quad 0 < p < 1,$$

   and N has the Poisson distribution with parameter $\lambda(>0)$. Find $E(S_N)$ and $V(S_N)$.

   (c) What is the actual distribution of S_N in part (b)?

2. Consider a bivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2)$ and dispersion matrix $\begin{pmatrix} 2.25 & 1.68 \\ 1.68 & 4.00 \end{pmatrix}$. Also, let $(\overline{X}_n, \overline{Y}_n)$ be the mean vector of a sample of size n from this distribution.

(i) Show that for testing

$$H_0: \theta = \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1: \theta > 0$$

the UMP test based on the set of n differences $(X_i - Y_i\,,\ 1 \le i \le n)$ consists in rejecting H_0 if $\overline{X}_n - \overline{Y}_n \geq c$ where $P\{\overline{X}_n - \overline{Y}_n \geq c \mid H_0\} = \alpha$, the level of significance.

(ii) Find an expression for the power function of the UMP test for $\theta = \theta_1 (>0)$.

(iii) Determine n such that for $\theta_1 = 0.5$ and $\alpha = 0.05$, the power is equal to 0.95.

(iv) For the same n and (α, θ_1), determine the power of the classical sign-test based on the signs of the difference of the two variates for the n observations (vectors).

3. Suppose that a set of three measurements (Y, X_1, X_2) is available on each experimental unit in a study.

(a) Determine the values of (β_1, β_2) which minimize the variance of

$$Z = Y - \alpha - \beta_1 X_1 - \beta_2 X_2 ,$$

where you may assume that the dispersion matrix of (Y, X_1, X_2) (say, equal to $\underset{\sim}{\Sigma}$) is known. Does the solution depend on α?

(b) Given a sample (Y_i, X_{1i}, X_{2i}), $i = 1,\dots,n$ from a trivariate normal distribution, find out the maximum likelihood estimator of the conditional mean of Y given $X_1 = x_1$ and $X_2 = x_2$.

(c) Suppose that in (b), it is given that the mean vector of the trivariate normal distribution is equal to $\underset{\sim}{0}$. Does the solution remain the same?

4. (a) Let X be a random variable with a continuous distribution function $F(x)$, $-\infty < x < \infty$. Find the distribution function of the random variable $Y = F(X)$. What is this result usually called?

(b) Illustrate the use of this result in simulating a random sample from a simple exponential distribution.

(c) Let $X_1,\dots,X_n$ be n independent random variables with continuous distribution functions $F_1(x),\dots,F_n(x)$, all defined on $(-\infty,\infty)$. Using the result in (a) show that each of $U_n = -2\sum_{i=1}^n \log F_i(X_i)$ and $V_n = -2\sum_{i=1}^n \log\{1-F_i(X_i)\}$ is distributed as chi-square with $2n$ degrees of freedom.

5. One way of dealing with the problem of non-response in a survey is to make efforts to collect information from a sub-sample of the units not responding in the first attempt. In a mail enquiry, n units were selected from a population of N using simple random sampling without replacement (SRSWOR). Of these, n_1 units responded. From the $n_2(=n-n_1)$ non-responding units, r_2 were selected with SRSWOR and the required information was obtained by personal interview.

(a) Let $n_2 = \nu r_2$, $\nu(>1)$ fixed in advance, N_1 be the number of units in the population that would have responded in the first attempt, $N_2 = N - N_1$ and $W_j = N_j/N$, $j = 1,2$. Show that

$$E(n_1/N_1) = \nu E(r_2/N_2) .$$

(b) Let y be the variable under study, $\bar{y}_1$ and $\bar{y}_2'$ be the sample means based on the n_1 units responding in the 1st attempt and on the sub-sample of size r_2 units respectively. Show that

$$\bar{y}' = (n_1\bar{y}_1 + n_2\bar{y}_2')/n$$

is an unbiased estimator of the population mean $\bar{Y}$.

(c) Show that

$$V(\bar{y}') = \frac{N-n}{N}\frac{S^2}{n} + \frac{(\nu-1)}{n} W_2 S_2^2 ,$$

where S^2 is the variance of the whole population and S_2^2 is the variance within the non-response stratum.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9:00 a.m. - 1:00 p.m., Saturday, January 22, 1977)

1. Suppose that after suitable observation, it is decided that the number (N) of auto accidents at a given intersection this year has a Poisson distribution with parameter λ. A typical accident involves one or more cars. Let X be a random variable denoting the number of autos involved in an accident. We assume that the numbers for different accidents are independent and identically distributed with some distribution.

   (a) Derive the distribution of T, the total number of autos involved in (N) accidents at the intersection this year. Name the resulting distribution.

   (b) Find the generating function of T.

   (c) If we are given the rule of thumb that, in practice, only one, two or three autos are involved in a given accident with corresponding frequencies in the ratio 5:3:2, what is the mean of T?

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter θ, i.e.,

$$f(x;\theta) = (x!)^{-1}\theta^x e^{-\theta}$$

   with $\theta > 0$ and $x = 0,1,2,\ldots$.

   Let $\Pr(X \leq 1) = (1+\theta)e^{-\theta} = \lambda$.

   (a) Show that $Y = \sum_1^n X_i$ is a complete sufficient statistic for θ.

   (b) Let $u(X_1) = 1$ for $X_1 \leq 1$ and $u(X_1) = 0$ for $X_1 > 1$. Show that $u(X_1)$ is unbiased for λ.

   (c) Obtain the minimum variance unbiased estimate for λ.

   (d) Show that the statistic obtained in (c) is indeed unbiased for λ.

3. In a certain region a 15% random sample of hotels is taken and classified according to whether or not every bedroom has a telephone. The results are:

Total Number of Bedrooms	Hotels in Region	Hotels in Sample	Sample Hotels With Full Number of Telephones
1-5	400	80	2
6-20	1500	200	50
21-50	600	100	45
51-100	300	50	35
101+	200	20	18
Total	3000	450	150

Compute the appropriate unbiased estimates of the total number of hotels with telephones in every room and the corresponding variance, assuming the data arose from

(a) a simple random sample without the knowledge of the strata,

(b) a stratified random sample as indicated in the table.

(c) Find the relative efficiency of the two methods of selection.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the following probability density function:

$$f(x;\theta_1,\theta_2) = \theta_2^{-1} e^{-(x-\theta_1)/\theta_2}$$

where $\theta_1 < x < \infty$, $-\infty < \theta_1 < \infty$, and $0 < \theta_2 < \infty$.

(a) Determine the most powerful test of size α for testing

$$H_0\colon \theta_2 = \theta_2' \text{ , with } \theta_1 \text{ known}$$

against

$$H_1\colon \theta_2 = \theta_2'' < \theta_2' \text{ , with } \theta_1 \text{ known .}$$

(b) Is the test obtained in (a) the uniformly most power test of size α for testing

$$H_0\colon \theta_2 = \theta_2' \text{ , with } \theta_1 \text{ known}$$

against

$$H_1\colon \theta_2 < \theta_2' \text{ , with } \theta_1 \text{ known ?}$$

(c) Obtain the likelihood ratio test for testing

$$H_0\colon \theta_1 = \theta_1' \text{ , with } \theta_2 \text{ unknown}$$

against

$$H_1\colon \theta_1 > \theta_1' \text{ , with } \theta_2 \text{ unknown .}$$

(d) Show that the test in (c) may be based on a statistic which has an F distribution under the null hypothesis.

5. A wrist-watch manufacturing concern intended to study the effect of humidity on the rusting of a particular metal they used for the watches. Two countries A (a tropical one) and B (a temperate one) were chosen and within each country five different places with varying degrees of humidity were selected. For each place, ten specimens of the metal were exposed to the normal environment for a year. For the i-th country, j-th place and k-th specimen, let Y_{ijk} denote the amount of rusting, and let t_{ij} be the average amount of humidity (during the experimental year) in the j-th place of the i-th country. Note that $i = A,B$; $j = 1,\ldots,5$; $k = 1,\ldots 10$.

It is conceived that for each country, a simple linear regression of Y on t describes the model adequately. The concern wants to know:

(a) Whether the rate of growth of rusting with humidiy is the same for both the countries, and

(b) Whether the typical rusting at 0 amount of humidity is the same for both the countries.

(i) State your assumptions carefully.

(ii) Provide suitable statistical tests for both (a) and (b).

(iii) Comment on their merits and demerits.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

SPECIAL OFFERING OF PART I

(9:00 a.m. - 1:00 p.m., Wednesday, August 3, 1977)

1. Sampling

Simple random samples drawn from each of three strata give data as follows:

Stratum I	Stratum II	Stratum III
10	21	30
12	23	31
13	25	28
17	23	40
8	20	38
9	26	37
20		37
11		33
13		
12		

(a) Estimate the population mean, and the variance of your estimate, under the assumption that the sample shown was drawn under proportional allocation.

(b) Estimate the population mean and the variance of your estimate, if the true stratum weights are equal.

(c) Assume that combined sample above was drawn by simple random sampling without stratification. Estimate the population mean and its variance.

(d) Now assume, under the conditions of (a), that data are available on a concomitant variable x, such that for each of the y values observed above, the corresponding x value is $(y-5)$. Suppose also that the actual totals for the x variable in the population are known to be 8, 12, and 20 in stratum I, II, III, respectively. Calculate the combined and separate ratio estimates of the population total Y. Which would you prefer in this case, and why?

2. Hypothesis Testing

Define

(a) Type I and Type II errors;

(b) Power of a test;

(c) Uniformly most power test;

(d) Unbiased test.

A box contains N items numbered from 1 to N, One item is selected at random and its number X observed. We wish to test the hypothesis $N=5$ at $\alpha=0.2$. The following 5 tests are thus available at the desired level: T_1 rejects H_0 if $X=1$ or $X>5$; T_2 rejects H_0 if $X=2$ or $X>5$; etc.

(e) Calculate and sketch the power curve of each of the tests $T_1, T_2, \ldots, T_5$ against the alternatives $N = 1,2,3,4,5,6,7,\ldots$. Are these tests unbiased?

(f) Is there a uniformly most powerful test among $T_1, T_2, \ldots, T_5$? Justify your answer.

3. Estimation

The random variable X is said to have a Pareto distribution if $F(x) = P\{X \le x\}$ is given by

$$F(x) = \begin{cases} 0 , & x \le x_0 \\ 1 - (x_0/x)^a , & x > x_0 \end{cases}$$

where both x_0 and a are positive numbers.

(a) Derive the form of the probability density function $f(x)$ corresponding to $F(x)$.

(b) Find $E(X^k)$ for $k = 1,2$ and $V(X)$.

(c) Under what condition on a does $E(X^k)$ exist for $k = 1,2$.

(d) If a is known and if $X_1,\ldots,X_n$ are independent random variables having the same distribution $F(x)$, show that there exists a sufficient statistics for x_0.

(e) For a specified a, find a 95% confidence interval for x_0 based on the sufficient statistic.

4. Stochastic Processes

The special case of the general birth and death process in which the birth rate is $\lambda_n = \lambda n + \alpha$, λ and $\alpha > 0$, and the death rate is μn, $\mu > 0$, is called a linear growth process with immigration.

(a) Write down the transition probabilities corresponding to the above birth and death rates. (We assume time homogeneity.)

(b) Is this a Markov process? Show why or why not.

(c) Determine differential-difference equations satisfied by the transition probabilities in (a).

(d) Use (c) to determine the (time-dependent) mean.

(e) Give conditions for which limits of the expressions in (a) and (d) exist and write down the limits.

5. Experimental Design

The problem often arises in experimental design over the large area: how many sites and how many replications for each experiment should be used to obtain some 'optimal' solution under certain conditions. We will consider the problem as follows:

Let y_{ij} denote the yield of some crop observed in the i-th experiment (site) $(i = 1,2,\ldots,m)$ on the j-th plot, $(j = 1,2,\ldots,r)$. (We assume that number of sites is very large, and for practical purpose infinite.)

Let

$$y_{ij} = \mu + u_i + \varepsilon_{ij} \ ,$$

with

$$\varepsilon_{ij} \frown N(0,\sigma_1^2) \ , \quad u_i \frown N(0,\sigma_2^2) \ ,$$

$$\mathrm{cov}(\varepsilon_{ij},\varepsilon_{i'j'}) = 0 \ , \quad \mathrm{cov}(\varepsilon_{ij},u_i) = 0 \ .$$

(a) Let

$$\bar{y} = \frac{1}{rm} \sum_{i=1}^{m} \sum_{j=1}^{r} y_{ij}$$

denote the observed overall average yield of this crop. Find $\mathrm{var}(\bar{y})$.

(b) As costs of an experiment, let c_1 be the cost of obtaining each plot and c_2 be the cost of preparing and conducting a trial within a plot. Give an expression for the total cost C of an experiment with m plots and r trials within a plot.

Find the optimal allocation with respect to r and m, when it is desired:

(c) to minimize $\mathrm{var}(\bar{y})$ for given total cost, $C = C_0$;

(d) to minimize total cost C for given variance $\mathrm{var}(\bar{y}) = V_0$.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 a.m. - 1 p.m., Saturday, January 21, 1978)

1. Consider a continuous process, denoting the interarrival time between the (i-1)th and the i^{th} events by T_i for $i \geq 1$. Presume that the T_i are independent and identically distributed as an exponential, specifically

$$f(t;\theta) = \theta e^{-\theta t}$$

for $t \geq 0$ and $\theta > 0$.

a) Verify that the no memory waiting time assumption is satisfied, i.e., $Pr(T > t + u | T > u) = Pr(T > t)$ for every $t \geq 0$, $u \geq 0$.

b) Let $T_1, T_2, \ldots, T_n$ be the interarrival times for the n events observed between time zero and time t. Find the distribution of the total time W_n required to observe n events.

c) Let N_t be the number of events occurring in an interval of length t. Find the distribution of N_t.

d) Suppose that

$$f(t;\theta) = \theta^2 t e^{-\theta t} \quad \text{for} \quad t \geq 0 \quad \text{and} \quad \theta > 0 .$$

Verify whether the no memory waiting time assumption is satisfied or not.

2. Let a random variable $Y = \log X$ be normally distributed with expected value ξ and standard deviation σ.

   a) Obtain the expected value of X, $E(X) = \mu$, say, in terms of ξ and σ.

   b) Let $X_1, X_2, \ldots, X_n$ be a random sample from this distribution. Find the maximum likelihood estimator of μ ($\hat{\mu}$, say).

   c) Find the expected value and variance of $\hat{\mu}$ (in terms of ξ and σ).

3. For the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e ,$$

suppose that (Y_i, x_{1i}, x_{2i}), $i = 1,\ldots,n$ are given.

a) What are the interpretations of the parameters $(\beta_0, \beta_1, \beta_2)$?

b) Suppose that you want to test

$$H_0: \beta_0 = 0 \text{ vs } H_1: \beta_0 \neq 0$$

where β_1, β_2 are treated as nuisance parameters. Is H_0 a linear hypothesis?

c) State clearly the standard assumptions usually made in this context.

d) Construct the standard analysis of variance table for this testing problem.

e) What is the distribution of the test statistic when (i) H_0 holds and (ii) H_0 does not hold?

4.

a) Define each of the following: i) Simple random sampling, ii) Stratified random sampling, iii) Systematic sampling, iv) Cluster sampling, and v) Two stage sampling.

b) Consider the ratio method of estimation, and let us denote by

y_i and x_i the values of the primary and an ancillary characteristic under study for the i^{th} unit of the population $(i = 1,2,\ldots,N)$

$R = \frac{\bar{Y}}{\bar{X}}$ the ratio of the population mean of y to the population mean of x

and

$\hat{R} = \frac{\bar{y}}{\bar{x}}$ the corresponding ratio for a sample of size n.

Show that $E(\hat{R})$ is approximately equal to $R\left\{1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_y C_x\right)\right\}$, where C_x, C_y denote the coefficient, of variation of x and y respectively, and ρ denotes the correlation coefficient between x and y.

c) Suppose that a population is divided into two mutually exclusive classes with N_1 and N_2 members respectively, so that

$$N_1 = Np \ , \quad N_2 = Nq \quad \text{and} \quad N_1 + N_2 = n \ .$$

Assume that a simple random sample of size n is drawn without replacement from the N units and n_1 of the observations in the sample are in class 1 and n_2 in class 2. Define

$$\hat{R} = \frac{n_1}{n_2} \quad \text{and} \quad R = \frac{N_1}{N_2} \ .$$

Show that for large values of N, the relative bias in $\hat{R}$ is given by $\frac{1}{nq}$.

5.

a) For testing a simple null hypothesis vs a simple alternative, what is a most powerful test? Does it exist always?

b) What is a uniformly most powerful test?

c) Let $X_1,\dots,X_n$ be n independent random variables each having the normal distribution with mean θ and variance θ, where $\theta > 0$. Let

$$H_0: \theta = 1 \qquad H_1: \theta = 1.5 .$$

(i) What is the most powerful test for this problem?

(ii) Verify whether the test derived in (i) is uniformly most powerful when

(I) $H_0: \theta = 1$ vs $H_1: \theta > 1$,

(II) $H_0: \theta = 1$ vs $H_1: \theta \neq 1$.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

SPECIAL OFFERING OF PART I

(9 a.m. - 1 p.m., Saturday, August 26, 1978)

QUESTION 1

a) State the components of variance model for the one-way layout consisting of k samples of n observations each. Use the notations σ_B^2 and σ_W^2 for the between-sample and within-sample components of variance, respectively.

b) Using the notations SSB and SSW for the between-sample and within-sample sums of squares, respectively, set up and fill out an analysis of variance table for this one-way layout, with rows "between" and "within" and columns "DF", "SS", "MS", "EMS", and "VR".

c) Let $\theta = \sigma_B^2/\sigma_W^2$. Give formulas for the point estimator of θ and for a two-sided $100(1-\alpha)\%$ confidence interval on θ. Note that θ cannot be negative. How do your formulas take this fact into account?

d) State how to test the hypothesis that $\theta = \theta_0$ against the one-sided alternative $\theta > \theta_0$. What is the power of the test against the specific alternative $\theta = \theta_1 > \theta_0$?

e) Suppose $k = 3$, $n = 6$; the 3 sample means are $\bar{x}_1 = 5$, $\bar{x}_2 = 6$, $\bar{x}_3 = 7$; and $SSW = 1800$, $\alpha = .10$, $\theta_0 = 1$, $\theta_1 = 2$. Write down the analysis of variance table for these data; calculate the point of estimate and the confidence interval of (c); perform the test of (d) and evaluate its power.

NOTE: If X is distributed as F with 2 and m degrees of freedom, then

$$P\{X > c\} = \left(\frac{m/2}{c + m/2}\right)^{m/2} ,$$

QUESTION 2

Suppose that X is normally distributed with mean ξ and variance σ^2.

a) Let $-\infty < L < U < \infty$. Derive the moment generating function of the conditional distribution of X given $L \le X \le U$.

b) Let $Y = (X - \xi)/\sigma$ and $\phi(y)$ be the probability density function of y. Show that

$$(d/dy)\phi(y) = -y\phi(y)$$

$$(d^2/dy^2)\phi(y) = (y^2 - 1)\phi(y), \quad \forall -\infty < y < \infty .$$

c) Using (a) or (b), derive

$$E(X | L \le X \le U) \quad \text{and} \quad Var(X | L \le X \le U)$$

in terms of ϕ and the corresponding cdf Φ.

d) Verify that as $L \to -\infty$ and $U \to +\infty$, the expressions in (c) converge to ξ and σ^2, respectively.

QUESTION 3

a) What is a likelihood function? Define a maximum likelihood estimator (MLE) of a parameter. State the usual properties of a MLE of a parameter θ.

b) Suppose that X, the length of life of an electric tube, has the simple exponential density function

$$f_\theta(x) = \theta e^{-\theta x}, \quad 0 \le x < \infty \qquad (\theta > 0) .$$

Based on a sample of size n, n_t have been observed to have lifetime less than $t (> 0)$ while $n - n_t$ have lifetime $\ge t$.

(i) Find the MLE $\hat{\theta}_n$ of θ.
(ii) Verify whether $\hat{\theta}_n$ is unbiased for θ or not.
(iii) What is the asymptotic variance of $\hat{\theta}_n$?

QUESTION 4

a) Define

(i) Markov Process

(ii) Markov Chain

b) For $n = 1, 2, \ldots,$ let $X_n = U_1 + U_2 + \cdots + U_n$ where $\{U_n: n \geq 1\}$ is a sequence of indpendent random variables. We consider two cases

(i) each U_n normally distributed,

(ii) each U_n takes the values 0 and 1 with probability p and q respectively $(p + q = 1;\ 0 < p < 1)$.

State in each case whether or not the stochastic process $\{X_n: n \geq 1\}$ is a Markov Process or Markov Chain. Explain your reasoning.

c) Let $\{X_1, X_2, \ldots\}$ be a sequence of independent and identically distributed random (non-negative) variables with p.d.f. $f(x)$ and let R be an integer valued random variable with Probability$\{R = r\} = p_r$, the X's and R being independent. A finite renewal process is one in which events occur at

$X_1, \quad X_2 + X_2, \ldots, X_1 + \ldots + X_R$.

Calculate

(i) the mean and variance of the distribution of the time to $X_1 + \ldots + X_R$

(ii) the expected number of events in $(0, t)$.

QUESTION 5

a) Explain each of the following

(i) Simple random sampling with or without replacement;

(ii) Stratified random sampling;

(iii) Cluster sampling.

b) In a population of size N, let p be the proportion of individuals with a certain characteristic. Let

$$Y_i = \begin{cases} 1, & \text{if the ith individual has the characteristic;} \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, N$.

Suppose that a simple random sample of size n is drawn without replacement. Denote by $\bar{y}_n$ the sample arithmetic mean of the Y-values. Show that $E\bar{y}_n = p$. Find $\mathrm{Var}(\bar{y}_n)$.

c) Assume that in the population described in (b) the proportion of individuals possessing the characteristic varies by age and sex. Let N_{hk} and n_{hk} denote respectively the number of people in h-th age and k-th sex group and the size of a simple random sample (drawn without replacement) from the corresponding group $(h = 1, 2, \ldots, H,\ k = 1, 2)$. Use the following scoring system. Denote

$$Y_{ihk} = 1, \quad \text{if i-th individual in the h-th age and k-th sex group possesses the characteristic}$$

$$= 0, \quad \text{otherwise}$$

$(i = 1, 2, \ldots, n_{hk},\ h = 1, 2, \ldots, H,\ k = 1, 2)$.

Use the scores to obtain an unbiased estimate of p, the proportion of individuals possessing the characteristic in the population. Find the variance of your estimate.

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

SPECIAL OFFERING OF PART I

(9 a.m. - 1 p.m., Saturday, August 11, 1979)

Question 1

Conditional on θ fixed, X and Y are independent random variables each having a Poisson distribution with expected value θ. Also, suppose that θ follows a distribution with the probability density function

$$g(\theta) = \frac{1}{\Gamma(\alpha)} e^{-\theta} \theta^{\alpha-1}, \quad 0 < \theta < \infty ,$$

where α is a positive integer.

(a) Find the marginal distribution of X (and Y).

(b) Find the joint distribution of X and Y .

(c) Find the value of the (linear) correlation coefficient between X and Y .

(d) Find the regression function of Y on X, namely, $E(Y|X=x)$.

Question 2

Let $(X_{11}, X_{12}, \ldots, X_{1n_1})$ and $(X_{21}, X_{22}, \ldots, X_{2n_2})$ be two independent random samples from Poisson distribution with parameters μ_1 and μ_2, respectively, $(\mu_1 > 0,\ \mu_2 > 0)$.

(a) Construct the likelihood and derive the maximum likelihood estimator for μ_1.

(b) Suppose that $\mu_1 = \mu_2 = \mu$. Obtain the maximum likelihood estimator of μ, using data from the two independent samples.

(c) Construct the (extended) likelihood ratio criterion for testing $H_0\colon \mu_1 = \mu_2 = \mu$ against the alternative $H_1\colon \mu_1 \neq \mu_2$, using a suitable distributional approximation for the likelihood ratio criterion.

(d) Suppose that the following data are available:

$$n_1 = 50, \quad \bar{x}_1 = 2.70; \quad n_2 = 60, \quad \bar{x}_2 = 3.05 ,$$

where

$$\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}, \qquad i = 1, 2.$$

Apply the test you construct in (c) to these data, using significance level $\alpha = 0.05$. What is your decision in regard to H_0?

Question 3

(a) Write down the estimator for a population proportion in the case of a stratified random sample.

(b) Derive a computational formula that would enable you to compute the variance of your estimator.

(c) Suppose that a simple random sample of size n_h was selected from each stratum of size N_h, $h = 1,\ldots,H$, and m_h respondents were obtained from each stratum. Then, for each stratum $h = 1,\ldots,H$, a simple random sub-sample of size t_h was selected for intensive follow-up from the $(n_h - m_h)$ non-respondents. Suppose further that the sub-samples of size t_h all responded, $h = 1,\ldots,H$. Consider the following estimator $\hat{p}_i$ for a population proportion, p_i, in category i,

$$\hat{p}_i = \frac{1}{N} \sum_{h=1}^{H} N_h \left[\frac{m_h \dfrac{m_{hi}}{m_h} + (n_h - m_h)\dfrac{t_{hi}}{t_h}}{n_h} \right],$$

where $N = \sum_{h=1}^{H} N_h$, m_{hi} is the number of respondents in category i out of the m_h initial respondents, and t_{hi} is the number of respondents in category i out of the sub-sample of size t_h, for $h = 1,\ldots, H$.

Show that the variance of $\hat{p}_i$ may be estimated using the following computational formula

$$\mathrm{var}(\hat{p}_i) = \frac{1}{N^2} \sum_{h=1}^{H} \left[N_h \frac{(N_h - m_h)(m_h - m_{hi})m_{hi}}{n_h^2(m_h - 1)} + N_h^2 \frac{(n_h - m_h - t_h)(t_h - t_{hi})(n_h - m_h)t_{hi}}{n_h^2\, t_h^2(t_h - 1)} \right].$$

(d) Suppose now that in the actual survey, some of the sub-sample of non-respondents did not respond even with intensive follow-up causing a few of the t_h to be zero. How might you modify the computational formula in part(c) to deal with this situation?

Question 4

Consider a one-way analysis of variance (ANOVA) model with fixed effects. There are $k(\geq 2)$ treatments, indexed $i = 1,\ldots,k$ and $n(\geq 2)$ replicates for each treatment, indexed $j = 1,\ldots,n$. The response for the jth replicate of the ith treatment is denoted by Y_{ij}, $1 \leq j \leq n$, $1 \leq i \leq k$.

(a) Write down the usual model for this situation and any assumptions about the model and its parameters.

(b) In terms of the model parameters, write down the null and alternative hypotheses for testing whether there are any treatment effects.

(c) Provide a complete breakdown of the usual ANOVA table, including degrees of freedom, sum of squares, mean squares.

(d) State the test statistic used for the test in part (b) and its distribution under the null hypothesis of no treatment effects.

(e) Derive the expected values of the mean square due to error and the mean square due to treatment when the null hypothesis may not be true.

(f) Write the formula for the noncentrality parameter of the distribution of the test statistic in part (d). Of what use is this parameter?

Question 5

(a) Give definitions of the following terms (define any notation you use):

(i) persistent positive (or positive recurrent) state

(ii) persistent null (or null recurrent) state

(iii) trasient state

(iv) stationary distribution

(b) Suppose you are given a finite Markov chain with transition matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & .1 & 0 & 0 & .9 \\ .3 & .1 & .1 & .3 & .2 \\ .2 & .2 & .2 & .2 & .2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(i) Find the conditional probability of eventual absorption into state 0, given $X_0 = i$, for $i = 0, 1, 2, 3, 4$.

(ii) Find the expected waiting time for absorption into state 0 or state r, given $X_0 = i$, for $i = 0, 1, 2, 3, 4$.

BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9:30 a.m. - 1:30 p.m., Saturday, August 9, 1980)

INSTRUCTIONS

a) This is a *closed-book* examination.

b) The time limit is four hours.

c) Answer *any four* (but *only* four) of the five questions which follow.

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

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

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

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

Question 1

a) Clarify the confusion of a person who makes the following statement:

> "Stratified sampling and cluster sampling are *conceptually identical*. Both are types of designs in which a probability sample is selected from a population which is divided into groups of elements."

b) Briefly describe stratified (simple) random sampling.

c) Let

$$Y_{hi} = \begin{cases} 1 \text{ if the i-th element in the h-th stratum possesses some attribute} \\ 0 \text{ if otherwise.} \end{cases}$$

and suppose we wish to estimate the proportion (P) of elements in a population of size N which posses the attribute. The sampling design we use is proportionate stratified (simple random) sampling and our estimator of P takes the general form

$$p = \sum_h^H W_h p_h$$

where $W_h = N_h/N$, N_h is the total number of elements in the h-th stratum, and p_h is the proportion of n_h sample elements possessing the attribute in the h-th stratum. Show that the true variance of p is

$$\text{Var}(p) = \frac{\sum_h^H W_h P_h Q_h}{n} - \frac{\sum_h^H W_h P_h Q_h}{N}$$

where $P_h = 1 - Q_h$ is the proportion of all elements in the h-th stratum which possess the attribute.

d) Assuming that $N_h \doteq N_h - 1$ and $N \doteq N - 1$, show that the true variance of the estimator (p_o) of P from a simple random sample of size n taken from the same population can be expressed as

$$\text{Var}(p_o) = \text{Var}(p) + \frac{1-f}{n} \sum_h^H W_h (P_h - P)^2$$

where $f = n/N$.

Question 2

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 $\mathrm{Var}(\hat{\theta}_n)$.

(d) Compare $\mathrm{Var}(\hat{\theta}_n)$ with the Cramér-Rao bound and comment on the efficiency and sufficiency of $\hat{\theta}_n$.

Question 3

a) Suppose that the probability of an insect laying $r (\geq 0)$ eggs is $e^{-m} m^r / r!$ and that the probability of an egg developing is $p (0 < p < 1)$. Assuming mutual independence of the eggs, show that the probability of a total $k (\geq 0)$ survivors is $e^{-mp} (mp)^k / k!$.

b) Suppose that two insects of the same species are placed in two different incubators with different temperatures, and you suspect that the temperature may have some effect on the probability of an egg developing, though the distribution of the number of eggs layed should be the same. Let p_1 and p_2 be these probabilities for the two insects and let k_1 and k_2 be the actual number of survivors in the two incubators

(i) Find the joint distribution of (k_1, k_2)

(ii) Find the conditional distribution of k_1, given $k = k_1 + k_2$.

(iii) We wish to test the hypothesis $H_0: p_1 = p_2$ against $H_1: p_1 > p_2$. It would be easier to construct the test if we use the distribution defined in (ii). How in this case H_0 is expressed? Construct a test of size $\alpha (0 < \alpha < 1)$, given k_1, k_2.

Question 4

Suppose that a random variable Y is related to a non-stochastic factor x over the interval $-1 \le x \le 1$ by the relationship

$$\eta_x = E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 .$$

Let N pairs of observations (x_i, Y_i), $i = 1,\ldots,N$, be available, for which $\mathrm{Var}(Y_i) = \sigma^2 (< \infty)$, $\mathrm{Cov}(Y_i, Y_j) = 0$, $\forall\ i \ne j = 1,\ldots,N$, and let $\mu_k = N^{-1}\sum_{i=1}^{N} x_i^k$, $k = 1, 2, 3$. Assume further that $\mu_1 = \mu_3 = 0$. It is desired to *approximate* the function η_x over the interval $-1 \le x \le 1$ by a linear function

$$\hat{\eta}_x = b_0 + b_1 x, \quad -1 \le x \le 1,$$

where

$$b_0 = N^{-1}\sum_{i=1}^{N} Y_i \quad \text{and} \quad b_1 = \sum_{i=1}^{N} x_i Y_i \Big/ \sum_{i=1}^{N} x_i^2 .$$

(a) Express $E(b_0)$ and $E(b_1)$ in terms of β's and μ's .

(b) Express $V(b_0)$ and $V(b_1)$ in terms of σ^2, N, β's and μ's .

(c) Show that

$$\mathrm{Cov}(b_0, b_1) = 0 \ (\text{since } \mu_1 = 0).$$

(d) Show in general that the *integrated mean square error* $J = \int_{-1}^{1} E[(\hat{\eta}_x - \eta_x)^2]dx$ can be written as the sum of the *integrated variance* V and the *integrated squared bias* B, where $V = \int_{-1}^{1} \mathrm{Var}(\hat{\eta}_x)dx$ and $B = \int_{-1}^{1} [E(\hat{\eta}_x) - \eta_x]^2 dx$.

(e) Show that, for the particular situation described earlier, V is minimized when $\mu_2 = 1$.

Question 5

Let $X_1,\ldots,X_n$ be n independent and identically distributed random variables with the probability density function (pdf) $f(x)$ and distribution function $F(x)$. Let $Z_1 \le \cdots \le Z_n$ be the order statistics corresponding to $X_1,\ldots,X_n$.

(a) For an arbitrary $k (1 \le k \le n-1)$, write down the joint pdf of $Z_1,\ldots,Z_{k+1}$

(b) Hence or otherwise, obtain the conditional pdf of Z_{k+1} given $Z_1,\ldots,Z_k$.

(c) Show that (b) depends on $Z_1,\ldots,Z_k$ only through Z_k.

(d) Define a Markov process.

(e) Define a Markov chain.

(f) Is the process $\{Z_1,\ldots,Z_n\}$ a Markov process? Is it a Markov chain? Does it have independent increments? Does it have homogeneous (stationary) increments.