MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0360664 9

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

ORNL 1582
Physics

A TWO SAMPLE PROCEDURE FOR LINEAR DISCRIMINATION IN NORMAL SAMPLES

Jack Moshman

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

LIBRARY LOAN COPY

DO NOT TRANSFER TO ANOTHER PERSON

If you wish someone else to see this document, send in name with document and the library will arrange a loan.

OAK RIDGE NATIONAL LABORATORY
OPERATED BY
CARBIDE AND CARBON CHEMICALS COMPANY
A DIVISION OF UNION CARBIDE AND CARBON CORPORATION
UCC
POST OFFICE BOX P
OAK RIDGE, TENNESSEE

ERRATA

p. 28, line 11--for "$\gamma \left[1 - \text{Prob } B > 0\right]$" read "$\gamma \left[1 - \text{Prob} \left\{B > 0\right\}\right]$".

p. 32, line 5(bottom)--for "where q denotes" read "where $\left[q\right]$ denotes".

p. 42, line 6--for "summing (3.25) over" read "summing (3.28) over".

ORNL 1582

This document consists of 93 pages.
Copy 7 of 281 copies. Series A.

Contract No. W-7405-eng-26

MATHEMATICS PANEL

A TWO SAMPLE PROCEDURE FOR LINEAR DISCRIMINATION IN NORMAL SAMPLES

Jack Moshman

DATE ISSUED

OAK RIDGE NATIONAL LABORATORY
Operated by
CARBIDE AND CARBON CHEMICALS COMPANY
A Division of Union Carbide and Carbon Corporation
Post Office Box P
Oak Ridge, Tennessee

ORNL 1582
Physics

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Engineering Library
7-11. Central Files
12. Central Files, ORNL R.C.
13. C. E. Larson
14. W. B. Humes (K-25)
15. L. B. Emlet (Y-12)
16. A. M. Weinberg
17. E. H. Taylor
18. E. D. Shipley
19. A. S. Householder
20. F. C. VonderLage
21. C. P. Keim
22. R. C. Briant
23. J. A. Swartout
24. J. A. Swartout
25. S. C. Lind
26. F. L. Culler
27. R. S. Livingston
28. A. H. Snell
29. A. Hollaender
30. M. T. Kelley
31. J. H. Frye, Jr.
32. G. H. Clewett
33. K. Z. Morgan
34. J. S. Felton
35. C. S. Harrill
36. C. E. Winters
37. D. W. Cardwell
38. E. M. King
39. Lewis Nelson
40. D. D. Cowen
41. J. Moshman
42. M. J. Skinner

EXTERNAL DISTRIBUTION

43. D. B. Benedict, C&CCC, South Charleston
44. R. F. Bacher, California Institute of Technology
45-281. Given distribution as shown in TID-4500 under Physics category

DISTRIBUTION PAGE TO BE REMOVED IF REPORT IS GIVEN PUBLIC DISTRIBUTION

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	Discrimination in linear normal regression	1
	General method for the construction of confidence intervals	3
	Purpose of this thesis	7
II.	THE ONE SAMPLE PROCEDURE	10
	Introduction	10
	A graphical construction of a confidence interval for x_o	10
	An analytical determination of a confidence interval for x_o	22
	The evaluation of the confidence coefficient	28
III.	A TWO SAMPLE PROCEDURE	31
	General description	31
	Confidence intervals in the two sample procedure	39
	The choice of z	41
	The choice of n_o	43
IV.	MULTIPLE LINEAR NORMAL REGRESSION SINGLE SAMPLE	50
	Introduction	50
	Discrimination in multiple, linear normal regression	50
	A geometric construction	54

CHAPTER PAGE

IV. (continued)

The analytic determination of a joint confidence region 58

V. MULTIPLE LINEAR NORMAL REGRESSION TWO SAMPLE 67

Introduction 67

The non-central F distribution 67

General description 68

The construction of a confidence region 73

The choice of z and n_o 77

VI. SUMMARY AND EXAMPLE 80

Summary 80

Example 81

BIBLIOGRAPHY 87

LIST OF FIGURES

FIGURE PAGE

1. Graphical Determination of Confidence Limits 8
2. Confidence Limits for x_o under Case (1a) 13
3. Confidence Limits for x_o under Case (1b) 15
4. Confidence Limits for x_o under Case (2a) 16
5. Confidence Limits for x_o under Case (2b) 18
6. Confidence Limits for x_o under Case (3a) 20
7. Confidence Limits for x_o under Case (3b) 21
8. Power Function of Test of Hypothesis that $\beta = 0$ 30
9. Expected Value, $E(n)$, and 95 Per Cent Point, $n_{.95}$, of Distribution of n as a Function of n_o . .48

CHAPTER I

INTRODUCTION

Discrimination in Linear Normal Regression

Let y denote a normally distributed, random variable with constant, unknown variance σ^2 and whose mean is an unknown linear function, $E(y) = \alpha^* + \beta^* x$, of a continuous, non-stochastic variable x. The probability density function $f(y; x)$ has the form

$$(1.1) \qquad f(y; x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y - \alpha^* - \beta^* x)^2}{2\sigma^2}\right].$$

The following problem is frequently encountered in numerous applications. A series of observations y_{oj} $(j = 1, 2, \ldots, m)$ is taken at a single, unknown value x_o of x. From these observations, it is desired that x_o be estimated. It is obvious that the solution of this problem requires the existence of some prior information about the values of α^*, β^* and σ^2.

It is well known [1, pp. 549-550] that unbiased estimates of the three unknown parameters may be obtained as follows: Make observations (x_i, y_i), $(i = 1, 2, \ldots, n)$ on y at n known values x_i of x. The x_i need not all be distinct, but at least two distinct values are required. Let $\bar{x} = \sum_{i=1}^{n} x_i/n$ and define new parameters α and β by $\beta = \beta^*$ and $\alpha = \alpha^* + \beta^* \bar{x}$. Thus $E(y)$ assumes the form

$$E(y) = \alpha + \beta(x - \bar{x}) \quad , \tag{1.2}$$

for an arbitrary x . Estimate α , β and σ^2 by

$$a = \sum_{i=1}^{n} y_i/n \quad ,$$

$$b = \sum_{i=1}^{n} y_i(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad , \tag{1.3}$$

and

$$s^2 = \sum_{i=1}^{n} \left[y_i - a - b(x_i - x)\right]^2 / (n - 2) \quad ,$$

respectively. The expected value of y , (1.2), may then be estimated by

$$\text{Est}\left[E(y)\right] = a + b(x - \bar{x}) \quad . \tag{1.4}$$

Possessing estimates of the parameters and letting $\bar{y}_o = \sum_{j=1}^{m} y_{oj}/m$, one may estimate the unknown value of x_o corresponding to the m observations y_{oj} by solving the equation

$$\bar{y}_o = a + b\left[\text{Est}(x_o) - \bar{x}\right] \quad , \tag{1.5}$$

for $\text{Est}(x_o)$.

An estimate of the variability of $\text{Est}(x_o)$ is needed to obtain a measure of the reliability of the statistic. Since $\text{Est}(x_o)$ is a continuous, random variable, the probability of its assuming any

particular value is zero. Non-trivial probability statements may be made about $\mathrm{Est}(x_o)$ only if one considers the probability of it falling in some region, rather than at some specified point.

The Neyman-Pearson [9] concept of a <u>confidence interval</u> used to estimate the parameter of a distribution will be employed. A procedure for the determination of a confidence interval for x_o from the data of the type described above is known but apparently not clearly understood by all statisticians. If one, for example, interprets literally the procedure described in a well known textbook on mathematical statistics [8, p. 301] it is quite possible to arrive at a confidence interval with end points that are complex numbers.

It is the purpose of this thesis to clarify the apparent anomaly. Furthermore, the two-sample technique of Stein [11] will be applied to devise an efficient design of the experimental procedure leading to the determination of confidence intervals.

General Method for the Construction of Confidence Intervals

It will be convenient at this point to sketch the general procedure that will be used in the sequel for the construction of confidence intervals.

Let O_n be a sample of n observations taken from some population with a probability density function of known form, but with unknown parameters θ, $\xi_1, \xi_2, \ldots, \xi_k$ $(k \geqslant 0)$. The parameter θ is the one for which a confidence interval will be constructed. Let the vector ξ denote the set $\xi_1, \xi_2, \ldots, \xi_k$ of nuisance parameters. The probability density function will be denoted by

$f(u; \theta, \xi)$. If there exist two functions $L_1(O_n)$ and $L_2(O_n)$, which depend on the sample but not on θ or ξ , then $L_i(O_n)$ $(i = 1, 2)$ are random variables and the interval

$$(1.6) \qquad \delta(O_n): \quad L_1(O_n) \leqslant \theta \leqslant L_2(O_n) \quad ,$$

is a random interval. If, independent of the true values of θ and ξ , the probability is $(1 - \gamma)$ that the random interval $\delta(O_n)$ covers θ , then $\delta(O_n)$ is termed a <u>confidence interval</u> for θ and $(1 - \gamma)$ is the associated <u>confidence coefficient</u>. The values of the functions $L_i(O_n)$ $(i = 1, 2)$ are called the <u>confidence limits</u> [13, pp. 122-126] .

Instead of finding the functions $L_i(O_n)$ $(i = 1, 2)$ and their distributions directly, it is frequently more expeditious to proceed in the following indirect manner. Let $\zeta(O_n; \theta)$ be some random variable which is a function of O_n and θ and whose distribution $g(\zeta)$ is known and independent of θ and ξ . Let $\psi(O_n)$ be a random variable such that ζ is a monotone increasing function of ψ , but ζ may also be a function of some other random vector variable, say $\rho(O_n)$. Let the probability density function of ψ , given ρ , be $h_\rho(\psi; \theta, \xi)$ which will generally be dependent on θ and ξ . For the purpose of this paper it will be sufficient to assume that the random variables ψ and ρ are statistically independent. Let

Prob_ρ (A) be the conditional probability of A given ρ. If

$$(1.7)\qquad \begin{aligned} 1 - \gamma &= \int_{\zeta_1}^{\zeta_2} g(\zeta)d\zeta = \text{Prob}\left\{\zeta \leqslant \zeta_2\right\} - \text{Prob}\left\{\zeta \leqslant \zeta_1\right\} \\ &= E\left[\text{Prob}_\rho\left\{\zeta \leqslant \zeta_2\right\} - \text{Prob}_\rho\left\{\zeta \leqslant \zeta_1\right\}\right]^{1} , \end{aligned}$$

and $\psi = \psi_i$ when $\zeta = \zeta_h$ $(i, h = 1, 2;\ i \neq h)$, then, regardless of the true values of θ and ξ, $\text{Prob}_\rho\left\{\psi \leqslant \psi_i;\ \theta, \xi\right\} =$ $\text{Prob}_\rho\left\{\zeta \leqslant \zeta_h\right\}$ is a random variable dependent on ρ and, from (1.7),

$$(1.8)\qquad 1 - \gamma = E\left[\text{Prob}_\rho\left\{\psi \leqslant \psi_1;\ \theta, \xi\right\} - \text{Prob}_\rho\left\{\psi \leqslant \psi_2;\ \theta, \xi\right\}\right] .$$

As the notation implies, ψ_1 and ψ_2 are functions of θ. Assume these functions to be monotone increasing and to range over the interval of definition of ψ as θ ranges over its set of admissible values. If ρ is given and if, for a given ψ, $L_1(O_n)$ and $L_2(O_n)$ are the values of θ for which $\psi = \psi_1$ and $\psi = \psi_2$ respectively, then

$$(1.9)\qquad \text{Prob}_\rho\left\{\psi \leqslant \psi_1;\ \theta, \xi\right\} = \text{Prob}_\rho\left\{\theta \geqslant L_1(O_n);\ \theta, \xi\right\} ,$$

---

[1] The expectation of a conditional probability will be taken to be the expectation with respect to the conditional vector variable. If $F(\rho)$ is the cumulative distribution function of ρ, then

$$E\left[\text{Prob}_\rho(A)\right] = \int \text{Prob}_\rho(A)\ dF(\rho) ,$$

where the integral is taken over the domain of definition of ρ.

and

$$\text{(1.10)} \qquad \text{Prob}_\rho \left\{ \psi \leq \psi_2;\ \theta,\ \xi \right\} = \text{Prob}_\rho \left\{ \theta \geq L_2(0_n);\ \theta,\ \xi \right\} \quad ,$$

whatever the true values of θ and ξ may be. Substituting (1.9) and (1.10) in (1.8), one obtains

$$\text{(1.11)} \qquad \begin{aligned} 1 - \gamma &= E\Big[\text{Prob}_\rho \left\{ \theta \geq L_1(0_n);\ \theta,\ \xi \right\} \\ &\qquad - \text{Prob}_\rho \left\{ \theta \geq L_2(0_n);\ \theta,\ \xi \right\}\Big] \\ &= \text{Prob} \left\{ \theta \geq L_1(0_n);\ \theta,\ \xi \right\} - \text{Prob} \left\{ \theta \geq L_2(0_n);\ \theta,\ \xi \right\} \\ &= \text{Prob} \left\{ L_1(0_n) \leq \theta \leq L_2(0_n);\ \theta,\ \xi \right\} \quad , \end{aligned}$$

the desired result.

In the event that ψ_1 and ψ_2 are monotone decreasing functions of θ, then L_1 is associated with ψ_2 and L_2 with ψ_1 in the same manner as before. In this event the analogs of (1.9) and (1.10) are

$$\text{(1.12)} \qquad \text{Prob}_\rho \left\{ \psi \leq \psi_1;\ \theta,\ \xi \right\} = \text{Prob}_\rho \left\{ \theta \leq L_2(0_n);\ \theta,\ \xi \right\} \quad ,$$

and

$$\text{(1.13)} \qquad \text{Prob}_\rho \left\{ \psi \leq \psi_2;\ \theta,\ \xi \right\} = \text{Prob}_\rho \left\{ \theta \leq L_1(0_n);\ \theta,\ \xi \right\} \quad ,$$

respectively which, when substituted in (1.8), again give (1.11).

The relationship described above may be exhibited graphically as in Figure 1. The two curved lines are $\psi_1(\theta)$ and $\psi_2(\theta)$. If ψ^* is the sample value of ψ, one erects a horizontal line at ψ^* which intersects $\psi_1(\theta)$ at A and $\psi_2(\theta)$ at B. The vertical projection of $\overline{AB}$ on the θ axis is the confidence interval $\delta(0_n)$. If θ_o is the true value of θ, then $\theta_o \in \delta(0_n)$ if, and only if, $\psi_2(\theta_o) \leq \psi \leq \psi_1(\theta_o)$. The confidence coefficient is $(1 - \gamma)$.

The preceding discussion may be generalized to construct a <u>confidence region</u> for two or more parameters. Conditions of monotonicity apply to functions of several variables in a similar way.

Purpose of This Thesis

Assume the normal, linear regression model described above. Let x_i $(i = 1, 2, \ldots, n)$ be arbitrarily specified, subject to the condition that at least two values be distinct. Corresponding to each x_i one makes an observation on the random variable y obtaining values y_i $(i = 1, 2, \ldots, n)$. In addition, m further values of y, say y_{oj} $(j = 1, 2, \ldots, m; m \geq 1)$, are observed where it is known that each y_{oj} corresponds to the same but unknown value x_o of x. The joint distribution of the entire sample of the $(n + m)$ observations is

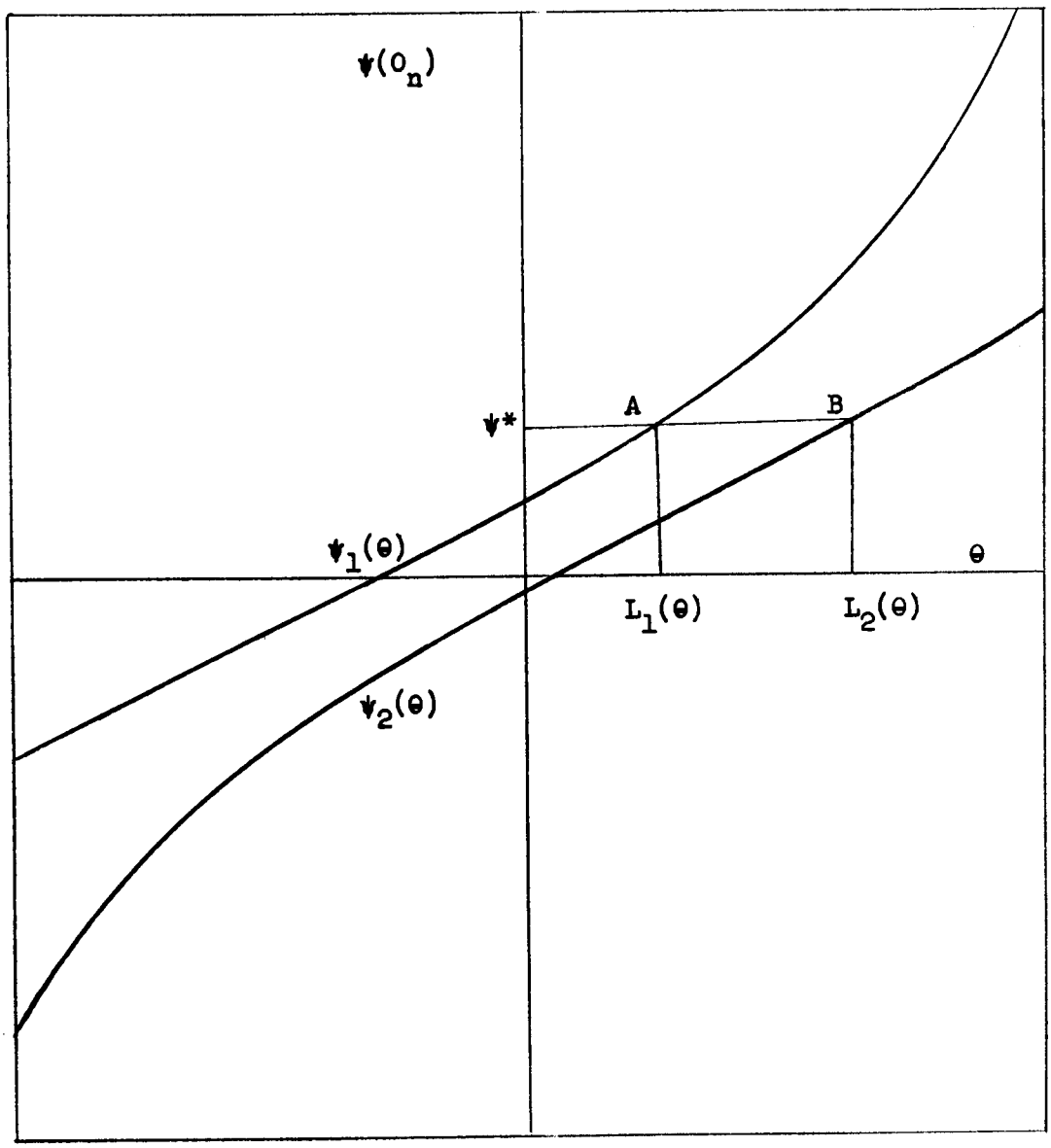

Figure 1

Graphical Determination of Confidence Limits

$$f(y_1, y_2, \ldots, y_n, y_{01}, y_{02}, \ldots, y_{om}; x_1, x_2, \ldots, x_n, x_0) =$$

(1.14)

$$\frac{1}{(2\pi\sigma^2)^{\frac{m+n}{2}}} \exp\left\{\frac{-\sum_{i=1}^{n}\left[y_i - \alpha - \beta(x_i-\bar{x})\right]^2 - \sum_{j=1}^{m}\left[y_{oj}-\alpha-\beta(x_o-\bar{x})\right]^2}{2\sigma^2}\right\}.$$

In (1.14) it is assumed that the x_i $(i = 1, 2, \ldots, n)$ are known, the y_i $(i = 1, 2, \ldots, n)$ and y_{oj} $(j = 1, 2, \ldots, m)$ are random variables and α, β, σ^2 and x_o are unknown parameters. This paper will describe a procedure for determining a confidence interval for x_o having the following properties:

1) The confidence interval will be unambiguous and not lead to any absurdities.

2) The sampling procedure will be efficient in that the sample size required will not be unnecessarily large to obtain an interval with the desired confidence coefficient.

3) The confidence interval will be independent of the nuisance parameters α, β and σ^2.

Chapter II will describe the usual one sample procedure. It will be seen that properties two and three are not fulfilled. Chapter III will outline the recommended two sample procedure. Chapters IV and V discuss the extension of the two sample method to higher dimensions. A summary and example will be found in Chapter VI.

CHAPTER II

THE ONE SAMPLE PROCEDURE

Introduction

This chapter contains an heuristic graphical description and a rigorous analytical discussion of the well-known one sample procedure for the construction of confidence limits for x_o .

A Graphical Construction of a Confidence Interval for x_o

The customary procedure used to determine a confidence interval for x_o is as follows. Compute the mean $\bar{y}_o = \sum_{j=1}^{m} y_{oj}/m$ of the m observations on y corresponding to the unknown x_o , and a , b and s^2 as defined in (1.3). Let $S = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $M = (\frac{1}{m} + \frac{1}{n})$ and

$$(2.1) \qquad u = \frac{\bar{y}_o - a - b(x_o - \bar{x})}{s\sqrt{M + \frac{(x_o - \bar{x})^2}{S}}} .$$

It is well known [4, p. 536] that u has Student's distribution with $(n - 2)$ degrees of freedom defined by

$$(2.2) \qquad g_{n-2}(u) = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi(n-2)}\,\Gamma(\frac{n-2}{2})} \left[1 + \frac{u^2}{n-2}\right]^{-\frac{n-1}{2}} .$$

Let $\nu = 1-\gamma/2$ and define $t_\nu(n - 2)$ to be the 100ν percentile of Student's distribution with $(n - 2)$ degrees of freedom, i.e.,

$$(2.3) \qquad \int_{-\infty}^{t_\nu(n-2)} g_{n-2}(u)du = 1 - \gamma/2 \quad .$$

Since $g_{n-2}(u)$ is symmetric about $u = 0$,

$$(2.4) \qquad \text{Prob}\left\{ |u| \leqslant t_\nu(n-2) \right\} = 1 - \gamma \quad .$$

Expanding and squaring the inequality within the brackets of (2.4) one has

$$(2.5) \qquad \frac{\left[\bar{y}_o - a - b(x_o - \bar{x})\right]^2}{s^2\left[M + \frac{(x_o - \bar{x})^2}{S}\right]} \leqslant t_\nu^2(n-2) \quad .$$

Defining

$$B = b^2 - \frac{t_\nu^2(n-2)\, s^2}{S} \quad ,$$

one may write (2.5) in the form

$$(2.6) \quad B(x_o - \bar{x})^2 - 2b(x_o - \bar{x})(\bar{y}_o - a) + (\bar{y}_o - a)^2 \leqslant M\, s^2\, t_\nu^2(n-2) \quad .$$

The equality in (2.6) defines a conic whose discriminant is

$$(2.7) \qquad 4\,b^2 - 4B = \frac{4\,s^2\,t_\nu^2(n-2)}{S} > 0 \quad .$$

The graph of the conic is an hyperbola. The inequality in (2.6)

specifies that the point $(x_o, \bar{y}_o)$ lies between the two branches of the graph. Thus, for a given value of $\bar{y}_o$, an interval, that may be of infinite extent, of values for x_o is specified. In many applications, the range of admissible values of x_o may be known to be finite, or semi-infinite. If this be the case, the confidence interval so obtained may be truncated to exclude inadmissible values of x_o.

The asymptotes of the hyperbola are obtained by setting the left-hand side of (2.6) equal to zero. The asymptotes are

$$\bar{y}_o - a = (x_o - \bar{x}) \left[b - \frac{s\, t_\nu(n-2)}{\sqrt{S}} \right], \tag{2.8}$$

and

$$\bar{y}_o - a = (x_o - \bar{x}) \left[b + \frac{s\, t_\nu(n-2)}{\sqrt{S}} \right]. \tag{2.9}$$

Several cases may be distinguished in terms of the slopes of the asymptotes:

<u>Case (1a)</u>. If

$$b > \frac{s\, t_\nu(n-2)}{\sqrt{S}}, \tag{2.10}$$

the slopes of both asymptotes are positive. Figure 2 illustrates the situation. Both branches of the hyperbola are continuous and monotonically increasing functions of x_o; they will be used to

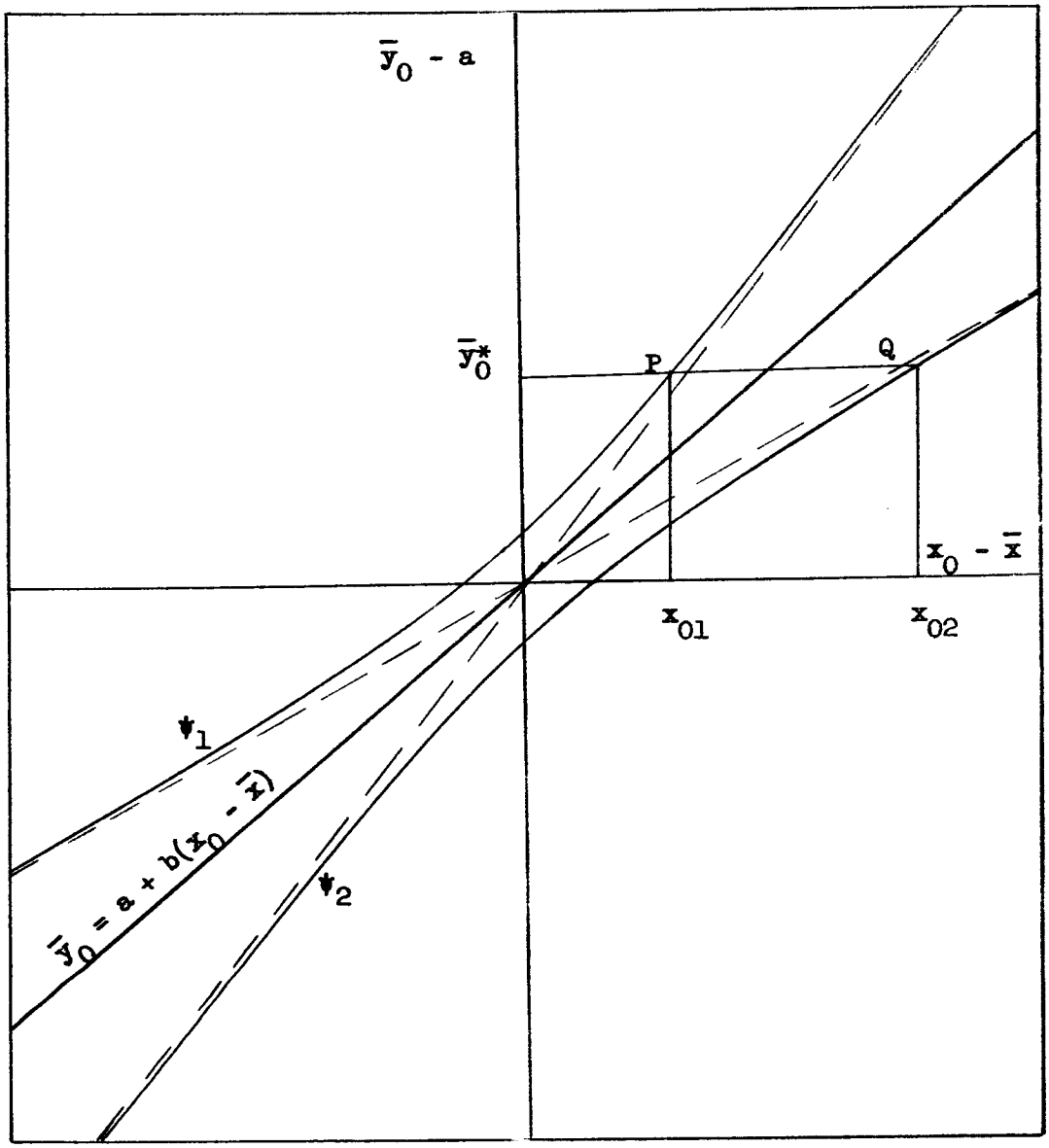

Figure 2

Confidence Limits for x_0 under Case (1a)

define the functions ψ_1 and ψ_2 discussed in Chapter I. Thus, if condition (2.10) holds, a conditional confidence interval for x_o may be obtained by intersecting the hyperbola with an horizontal line through the observed value, say $\bar{y}_o^*$. The projection $x_{01} \leqslant x_o \leqslant x_{02}$ of the line segment $\overline{PQ}$ is the desired interval. The conditional probability, under (2.10), that this interval covers the true x_o is $(1 - \gamma)$, since this interval is derived from the inequality (2.6), which in turn is just a restatement of the inequality in (2.4).

<u>Case (1b)</u>. If

$$b < -\frac{s\, t_\nu(n-2)}{\sqrt{S}} , \tag{2.11}$$

the slopes of both asymptotes are negative. Figure 3 depicts the construction of the confidence interval. The situation is entirely analogous to Case (1a) except for the reversal of the sign of the slopes of the asymptotes.

Case (1a) or (1b) results if $B > 0$.

<u>Case (2a)</u>. If

$$0 < b < \frac{s\, t_\nu(n-2)}{\sqrt{S}} , \tag{2.12}$$

the slopes of the asymptotes are of opposite sign as in Figure 4. The two branches are not admissible definitions of the functions ψ_1 and ψ_2 for they are not monotone functions of x_o. An horizontal line

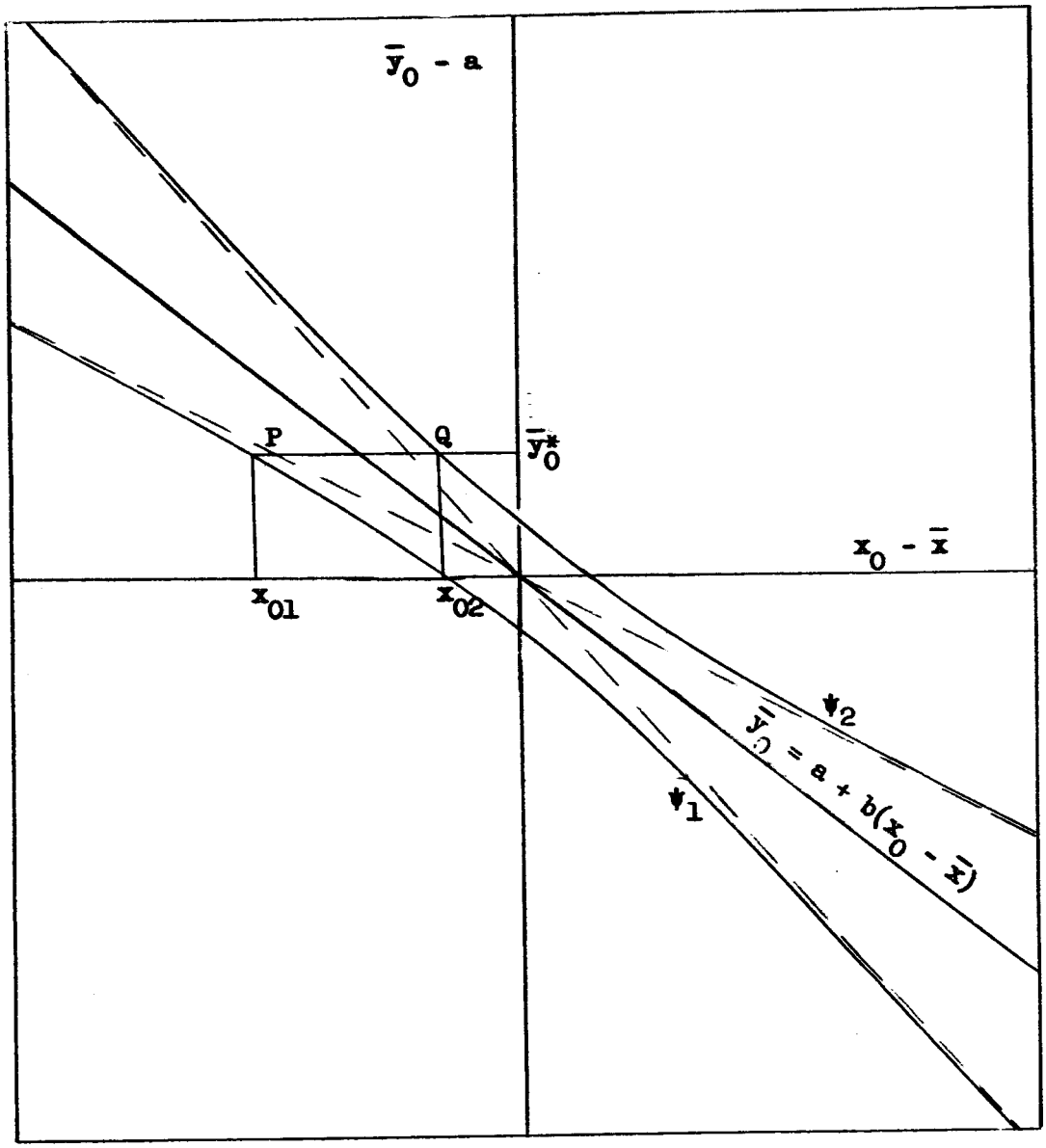

Figure 3

Confidence Limits for x_0 under Case (1b)

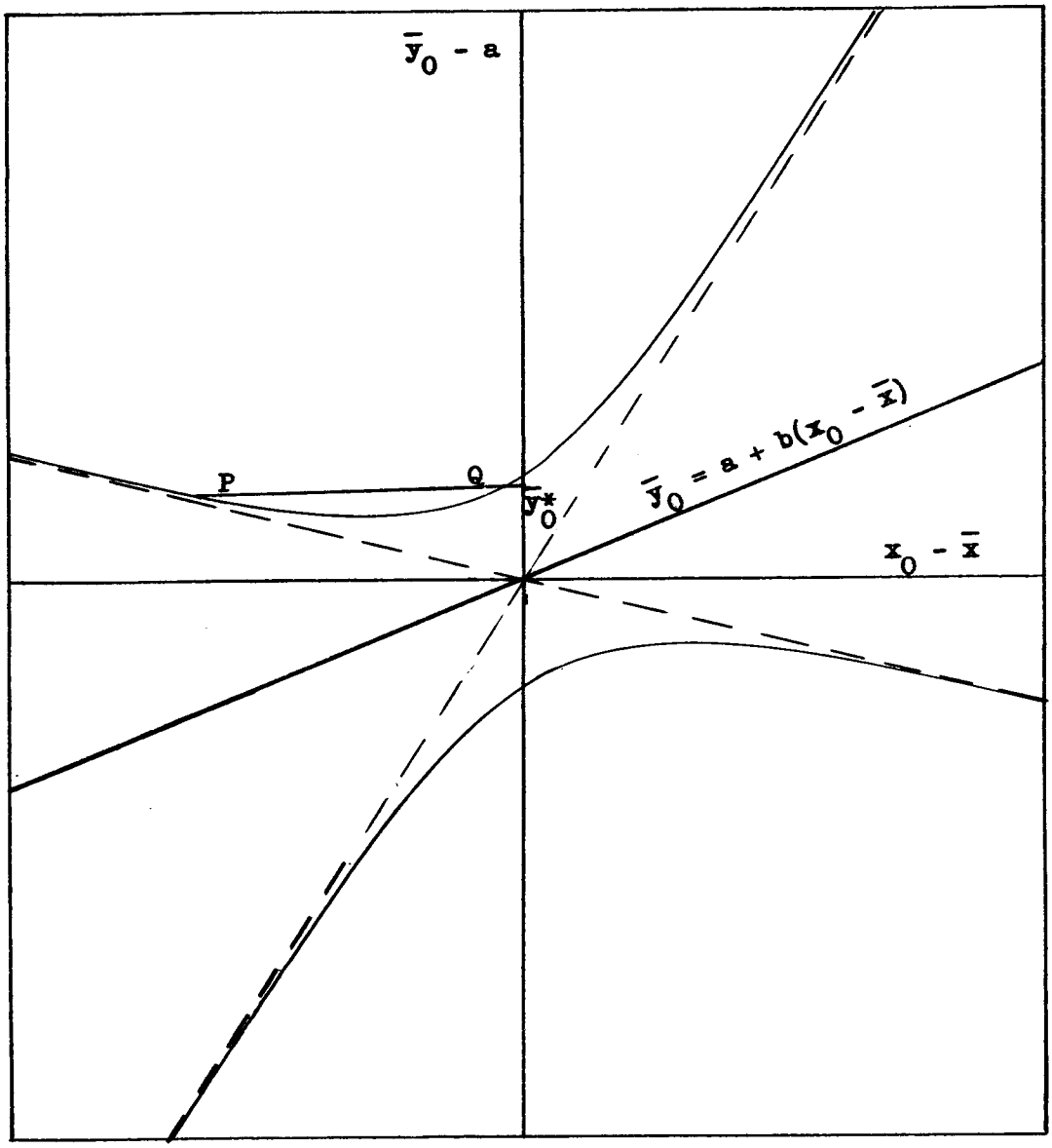

Figure 4

Confidence Limits for x_0 under Case (2a)

at $\bar{y}_o^*$ intersects the same branch of the hyperbola twice, at P and at Q . It will be convenient, and realistic, to define the conditional confidence interval for x_o to be the entire real line whenever (2.12) holds, although rigorously one does not have a conditional confidence interval in this case.

Case (2b). If

$$(2.13) \qquad 0 > b > -\frac{s\, t_\nu(n-2)}{\sqrt{S}} ,$$

the dual of Case (2a) is obtained. The same definition will be applied to construct a conditional confidence interval when (2.13) holds. Figure 5 illustrates this case.

Case (2a) or (2b) results if $B < 0$. One cannot strictly speak of the probability that the real line covers x_o . If $B < 0$, the sample actually does not provide enough evidence for the dependence of y on x .

Case (3a). In the event that

$$(2.14) \qquad b = \frac{s\, t_\nu(n-2)}{\sqrt{S}} ,$$

the slope of one asymptote is zero and that of the other is positive. The two branches of the hyperbola are monotone functions of x_o and may be taken to define ψ_1 and ψ_2 . If (2.14) holds, a conditional confidence interval for x_o may be obtained by intersecting the hyperbola with an horizontal line through the observed value $\bar{y}_o^*$.

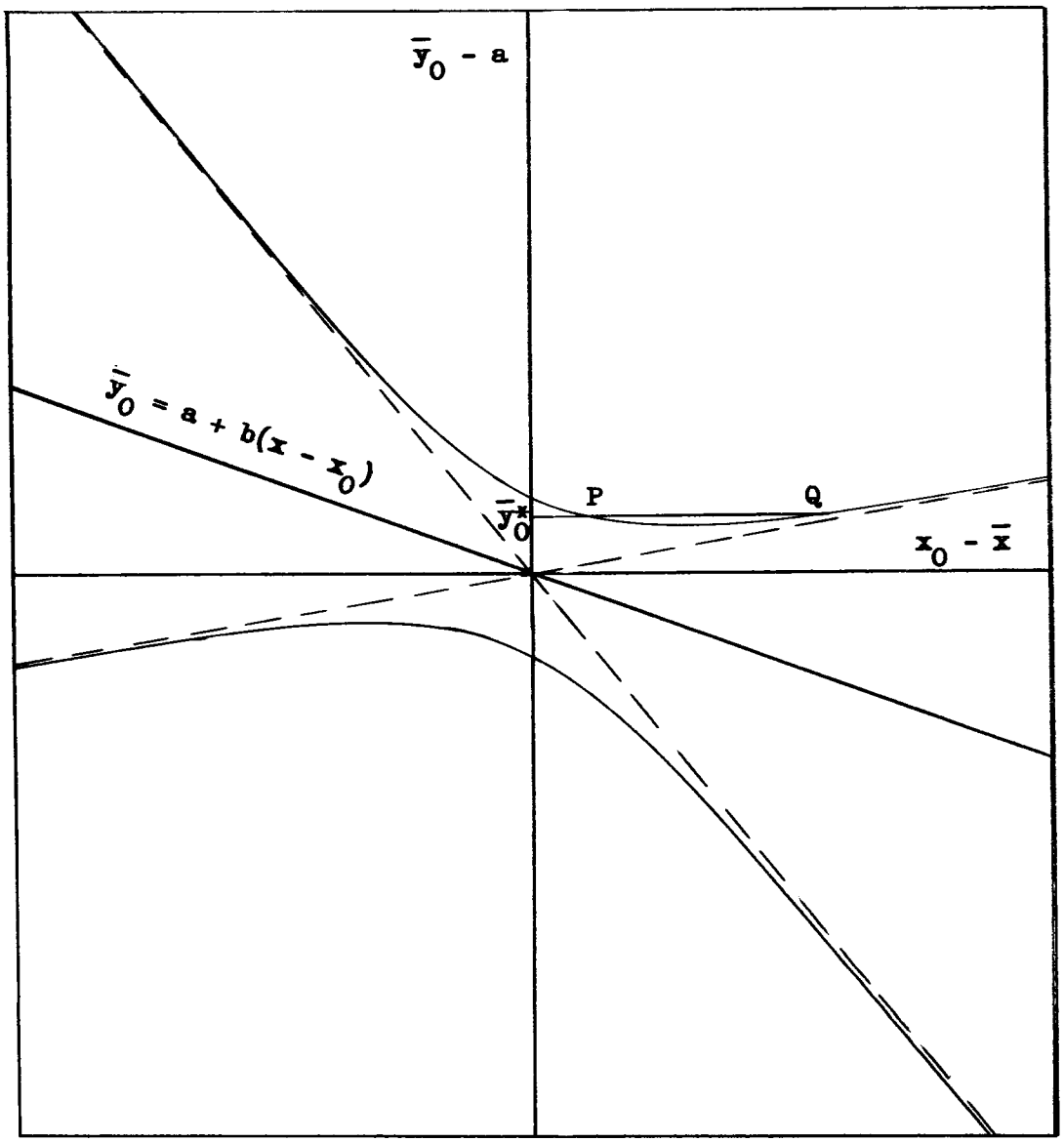

Figure 5

Confidence Limits for x_0 under Case (2b)

The line intersects but one branch if $\bar{y}_o^* \neq a$, and none otherwise. If $\bar{y}_o^* > a$, the conditional confidence interval is the projection, $x_{o1} \leqslant x_o$, of the half-line commencing at P . If $\bar{y}_o^* < a$, the conditional confidence interval is the projection, $x_o \leqslant x_{o2}$, of the half-line terminating at Q . If $\bar{y}_o^* = a$, the conditional confidence interval will be defined to be $-\infty < x_o < \infty$ in the same sense as in Case (2a). Case (3a) is illustrated in Figure 6.

<u>Case (3b)</u>. If

$$b = -\frac{s\, t_\nu(n-2)}{\sqrt{S}} \quad , \tag{2.15}$$

the slope of one asymptote is negative and that of the other is zero as in Figure 7. Conditional confidence intervals are constructed in a manner analogous to that described in Case (3a). If $\bar{y}_o^* > a$, the conditional confidence interval is the projection $x_o \leqslant x_{o2}$ of the half line terminating at P . If $\bar{y}_o^* < a$, the conditional confidence interval is the projection $x_{o1} \leqslant x_o$ of the half line commencing at Q . If $\bar{y}_o^* = 0$, the conditional confidence interval will be defined to be $-\infty < x_o < \infty$ in the same sense as Case (2b).

Case (3a) or (3b) results if $B = 0$. The confidence coefficient is not readily determinable, but since b is a continuous random variable, the probability of Case (3a) or (3b) occurring is zero. These cases are sketched for the sake of completeness and as a definition to serve in actual practice when (2.14) or (2.15) may hold

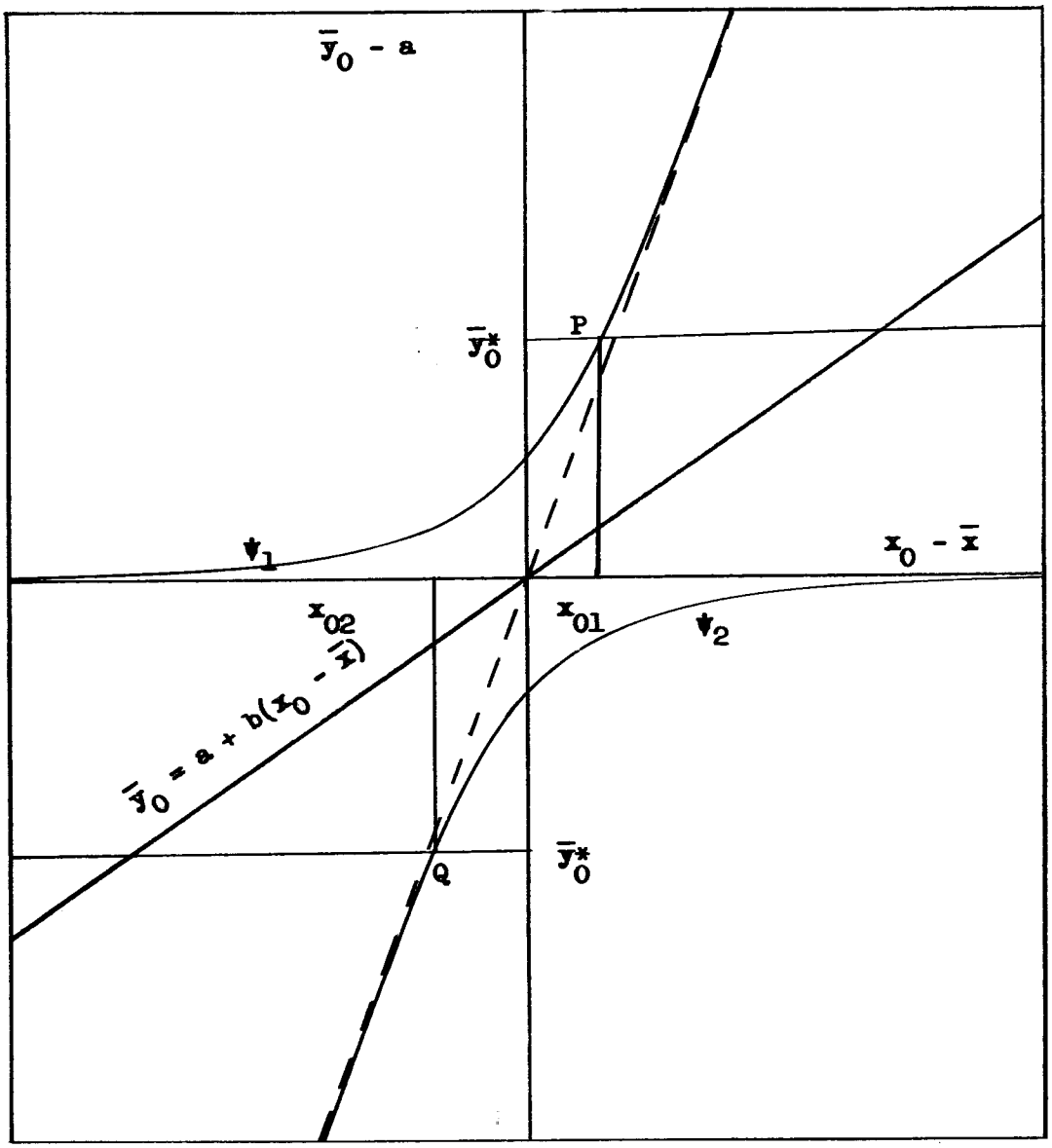

Figure 6

Confidence Limits for x_0 under Case (3a)

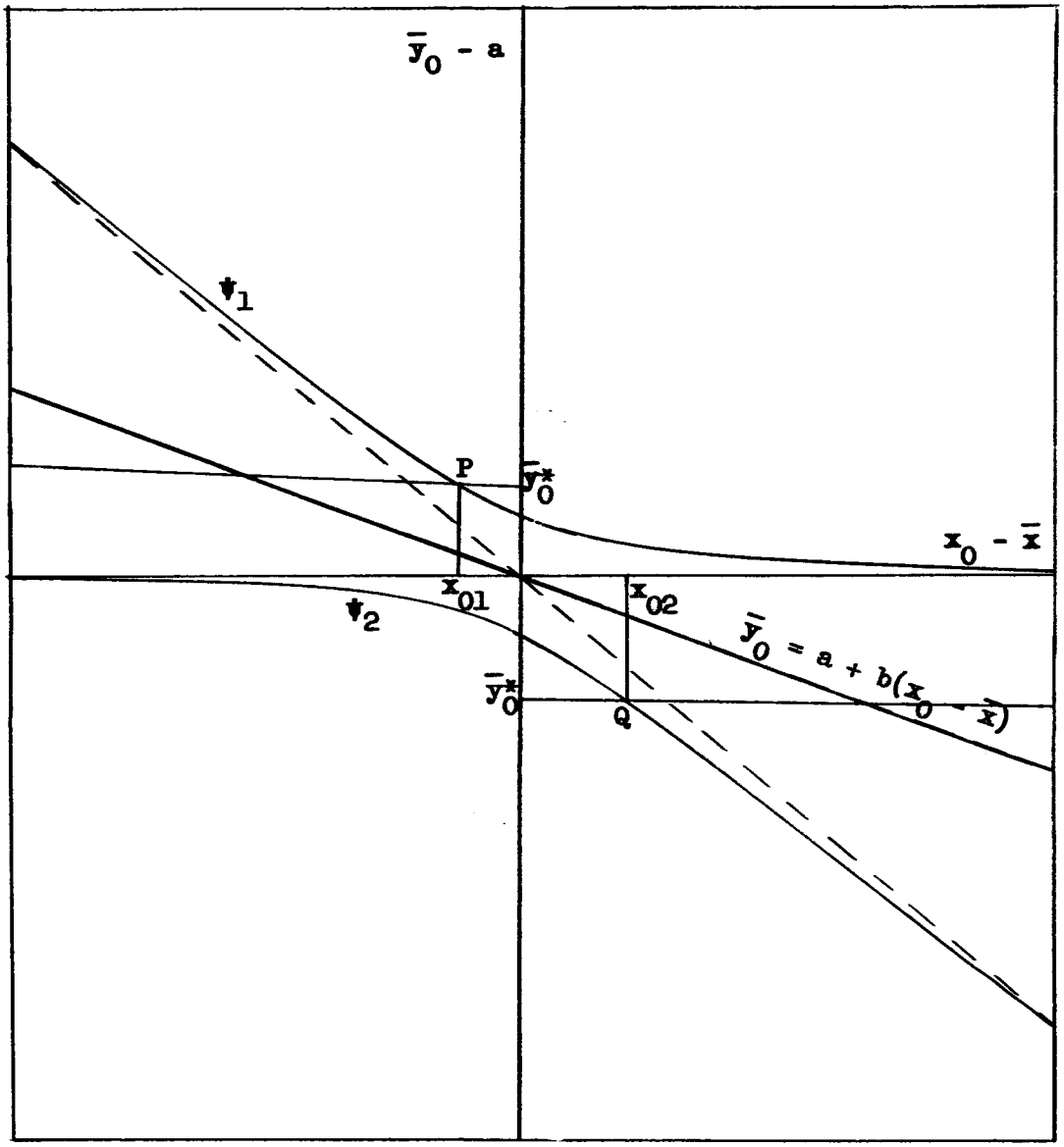

Figure 7

Confidence Limits for x_0 under Case (3b)

to within the number of significant digits used in the calculations.

An Analytical Determination of a Confidence Interval for x_o

The graphical construction described in the previous section is heuristic. Conditional confidence intervals for x_o are provided under various conditions on b. In order to establish rigorously, that the combined procedure furnished a rule for the determination of unconditional confidence intervals in the sense of Neyman and Pearson [9], the procedure will be developed analytically. A pair of random variables $L_i(O_n)$ $(i = 1, 2)$ that are functions of the sample O_n will be defined in such a manner that

$$\text{(2.16)} \qquad \text{Prob}\left\{ L_1(O_n) \leqslant x_o \leqslant L_2(O_n);\ x_o \right\} \geqslant 1 - \gamma ,$$

independent of the true value of x_o. Consideration will be given to the amount that the probability (2.16) exceeds $(1 - \gamma)$. The excess will be taken as a measure of the inefficiency of the procedure.

Definition. Let k be the observed value of the variable $(\bar{y}_o - a)$ and $n' = (n - 2)$. Define $L_i(O_n)$ $(i = 1, 2)$ as functions of the sample O_n by

$$\text{(2.17)} \quad L_1(O_n) = \begin{cases} \bar{x} + \dfrac{bk}{B} - \dfrac{s\, t_\nu(n')}{B} \sqrt{\dfrac{k^2}{S} + M B} & \text{if } B > 0 \\ \bar{x} + \dfrac{k}{2b} - \dfrac{b\, S\, M}{2k} & \text{if } B = 0 \text{ and } k > 0 \\ -\infty & \text{otherwise} , \end{cases}$$

and

$$(2.18)\quad L_2(0_n) = \begin{cases} \bar{x} + \dfrac{bk}{B} + \dfrac{s\, t_\nu(n')}{B}\sqrt{\dfrac{k^2}{S} + M\,B} & \text{if } B > 0 \\[2ex] \bar{x} + \dfrac{k}{2b} - \dfrac{b\,S\,M}{2k} & \text{if } B = 0 \text{ and } k < 0 \\[2ex] +\infty & \text{otherwise.} \end{cases}$$

These functions, so defined, agree with the limits set in the discussion of the graphical procedure above. Statement (2.16) will now be sharpened and proven in

<u>Theorem 2.1</u>. Let $L_i(0_n)$ $(i = 1, 2)$ be two functions defined by (2.17) and (2.18) respectively and let γ be a pre-assigned number such that $0 \leqslant \gamma \leqslant 1$. Define $\nu = 1 - \gamma/2$ and

$$B = b^2 - \frac{s^2\, t_\nu^2(n-2)}{S} .$$

Regardless of the true value of x_o,

$$(2.19)\quad \begin{aligned} &\text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} = \\ &(1-\gamma) + \gamma\left[1 - \text{Prob}\left\{B > 0\right\}\right] . \end{aligned}$$

<u>Proof</u>. By well-known rules of probability

$$\text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} =$$

$$\text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n),\ B > 0;\ x_o\right\}$$

$$(2.20) \quad + \text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n),\ B = 0;\ x_o\right\}$$

$$+ \text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n),\ B < 0;\ x_o\right\} .$$

Now, if $B < 0$, $L_1(0_n) = -\infty$ and $L_2(0_n) = \infty$ so that the inequality $L_1 \leqslant x_o \leqslant L_2$ imposes no restrictions on the parameter. Thus the third term in (2.20) is

$$(2.21) \quad \begin{aligned} &\text{Prob}\left\{-\infty < x_o < \infty,\ B < 0;\ x_o\right\} \\ &= \text{Prob}\left\{B < 0\right\} = (1 - \gamma + \gamma)\ \text{Prob}\left\{B < 0\right\}. \end{aligned}$$

Since the sample set on which $B = 0$ has zero probability,

$$(2.22) \quad \begin{aligned} &\text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n),\ B = 0;\ x_o\right\} \\ &= \text{Prob}_{B=0}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} \\ &\quad \cdot \text{Prob}\left\{B = 0\right\} = 0 . \end{aligned}$$

To evaluate the first term in (2.20), let $B > 0$ and consider

(2.1). The variable u has Student's distribution with $n' = (n - 2)$ degrees of freedom. Hence, if $\nu = 1 - \gamma/2$, then

$$1 - \gamma = \int_{-t_\nu(n')}^{t_\nu(n')} g_{n'}(u)\, du = \tag{2.23}$$

$$\text{Prob}\left\{u \leq t_\nu(n')\right\} - \text{Prob}\left\{u \leq -t_\nu(n')\right\} .$$

But u is a monotonic increasing function of $(\bar{y}_o - a)$, as well as a function of s and b. However $(\bar{y}_o - a)$ is statistically independent of s and b, so that, given the latter, one may write

$$1 - \gamma = E\left[\text{Prob}_{s,b}\left\{\bar{y}_o - a \leq b(x_o - \bar{x}) + s\, t_\nu(n') \sqrt{M + \frac{(x_o - \bar{x})^2}{S}}\, ;\, x_o\right\} - \right. \tag{2.24}$$

$$\left. \text{Prob}_{s,b}\left\{\bar{y}_o - a \leq b(x_o - \bar{x}) - s\, t_\nu(n') \sqrt{M + \frac{(x_o - \bar{x})^2}{S}}\, ;\, x_o\right\}\right] .$$

It may be seen that u plays the role of ζ and $(\bar{y}_o - a)$ that of ψ in the discussion in Chapter I. The random variables s and b correspond to the vector ρ. Furthermore the limits of ψ are

$$\psi_i(x_o) = b(x_o - \bar{x}) + (-1)^{i+1} s\, t_\nu(n') \sqrt{M + \frac{(x_o - \bar{x})^2}{S}} , \tag{2.25}$$

$$(i = 1, 2) .$$

Now

$$\frac{d\psi_i(x_o)}{d(x_o-\bar{x})} = b + (-1)^{i+1} \frac{s\, t_\nu(n')\ (x_o - \bar{x})}{S\sqrt{M + \frac{(x_o - \bar{x})^2}{S}}}$$

(2.26)

$$= b + (-1)^{i+1} \frac{s\, t_\nu(n')}{\sqrt{S}\sqrt{1 + \frac{M\,S}{(x_o - \bar{x})^2}}} \quad .$$

If $b > 0$, then

$$\frac{d\psi_2(x_o)}{d(x_o-\bar{x})} > b - \frac{s\, t_\nu(n')}{\sqrt{S}} > 0 \ , \tag{2.27}$$

since $B > 0$. Obviously $d\psi_1(x_o)/d(x_o - \bar{x}) > 0$. Thus if $b > 0$, $\psi_1(x_o)$ and $\psi_2(x_o)$ are monotone increasing functions of $(x_o - \bar{x})$ and hence of x_o. If $b < 0$, then $d\psi_2(x_o)/d(x_o - \bar{x}) < 0$ and

$$\frac{d\psi_1(x_o)}{d(x_o-\bar{x})} < b + \frac{s\, t_\nu(n')}{\sqrt{S}} < 0 \ , \tag{2.28}$$

because $B > 0$. Thus, if $b < 0$, $\psi_1(x_o)$ and $\psi_2(x_o)$ are also monotone decreasing functions of $(x_o - \bar{x})$ and hence of x_o.

The following argument will assume that $b > 0$. The modifications are obvious if $b < 0$. The functions $\psi_1(x_o)$ and $\psi_2(x_o)$ are similar to those depicted in Figure 2. Corresponding to

an observed value k of $(\bar{y}_o - a)$, $x_o = L_i(0_n)$ when $(\bar{y}_o - a) = \psi_i(x_o)$ $(i = 1, 2)$. Thus

$$
\begin{aligned}
1 - \gamma &= E\left[\mathrm{Prob}_{s,b}\left\{\bar{y}_o - a \leqslant \psi_1(x_o);\ x_o\right\} - \mathrm{Prob}_{s,b}\left\{\bar{y}_o - a \leqslant \psi_2(x_o);\ x_o\right\}\right] \\
&= E\left[\mathrm{Prob}_{s,b}\left\{x_o \geqslant L_1(0_n);\ x_o\right\} - \mathrm{Prob}_{s,b}\left\{x_o \geqslant L_2(0_n);\ x_o\right\}\right] \qquad (2.29) \\
&= E\left[\mathrm{Prob}_{s,b}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\}\right] \\
&= \mathrm{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} .
\end{aligned}
$$

Throughout the above discussion it was assumed that $B > 0$. Hence

$$
\mathrm{Prob}_{B>0}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} = 1 - \gamma , \qquad (2.30)
$$

and therefore

$$
\begin{aligned}
&\mathrm{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n),\ B > 0;\ x_o\right\} \\
&= \mathrm{Prob}_{B>0}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} \mathrm{Prob}\left\{B > 0\right\} \qquad (2.31) \\
&= (1 - \gamma)\ \mathrm{Prob}\left\{B > 0\right\} .
\end{aligned}
$$

Substituting (2.21), (2.22) and (2.31) into (2.20), one obtains the unconditional probability

$$\text{Prob}\left\{L_1(0_n) \leqslant x_o \leqslant L_2(0_n);\ x_o\right\} =$$

$$(1 - \gamma + \gamma)\ \text{Prob}\left\{B < 0\right\} + 0 + (1 - \gamma)\ \text{Prob}\left\{B > 0\right\}$$

(2.32)

$$= (1 - \gamma)\left[\text{Prob}\left\{B > 0\right\} + \text{Prob}\left\{B < 0\right\}\right] + \gamma\ \text{Prob}\left\{B \leqslant 0\right\}$$

$$= (1 - \gamma) + \gamma\left[1 - \text{Prob}\left\{B > 0\right\}\right] .$$

This concludes the proof of Theorem 2.1.

<u>The Evaluation of the Confidence Coefficient</u>

Theorem 2.1 proves that the usual analytic determination of a confidence interval for x_o does not result in the pre-assigned confidence coefficient but exceeds that amount by $\gamma\left[1 - \text{Prob } B > 0\right]$. An efficient procedure is one which will reduce the excess to zero. If the excess is greater than zero, part of the sample is, in a sense, being wasted in obtaining a confidence coefficient larger than that originally specified. The probability that $B > 0$ is equal to the <u>power</u> of a test of the hypothesis that $\beta = 0$.

The <u>power</u> of a test of hypothesis in the Neyman-Pearson [9] theory is the probability of rejecting the hypothesis. It would be desirable that the power be low when the hypothesis is true and large otherwise. The power function of any non-trivial one sample test of

the hypothesis that $\beta = 0$ depends on the true value of β, as it intuitively should, but also on the true value of σ^2 [2]. Thus one would have to know the true variance in order to control the excess over $(1 - \gamma)$.

The power of the one sample test of $\beta = 0$ is evaluated by means of the non-central t distribution [7] and is actually a function of β/σ. Figure 8 depicts the curve of a typical power function when $\gamma = .05$. If the power is large, then the analytic procedure is efficient; if the power is small, the procedure is inefficient. If the power is small, then β does not significantly differ from zero and the dependence of y on x is not demonstrated. Hence, knowledge of $\bar{y}_o$ should not give one information about its associated x_o.

Another interpretation of the power is that the confidence interval is finite if, and only if, $B > 0$. Thus, a relatively inefficient procedure results when the a priori probability is large that β does not differ significantly from zero. Since this probability requires information about both β and σ^2, one cannot generally control the excess over $(1 - \gamma)$ in advance.

The following chapter describes a two-sample procedure that has the advantages of making the power independent of the variance and of further reducing the excess in the confidence coefficient to any arbitrary quantity δ, in the range $0 < \delta < \gamma$, whenever β is outside some pre-specified zone of indifference about zero.

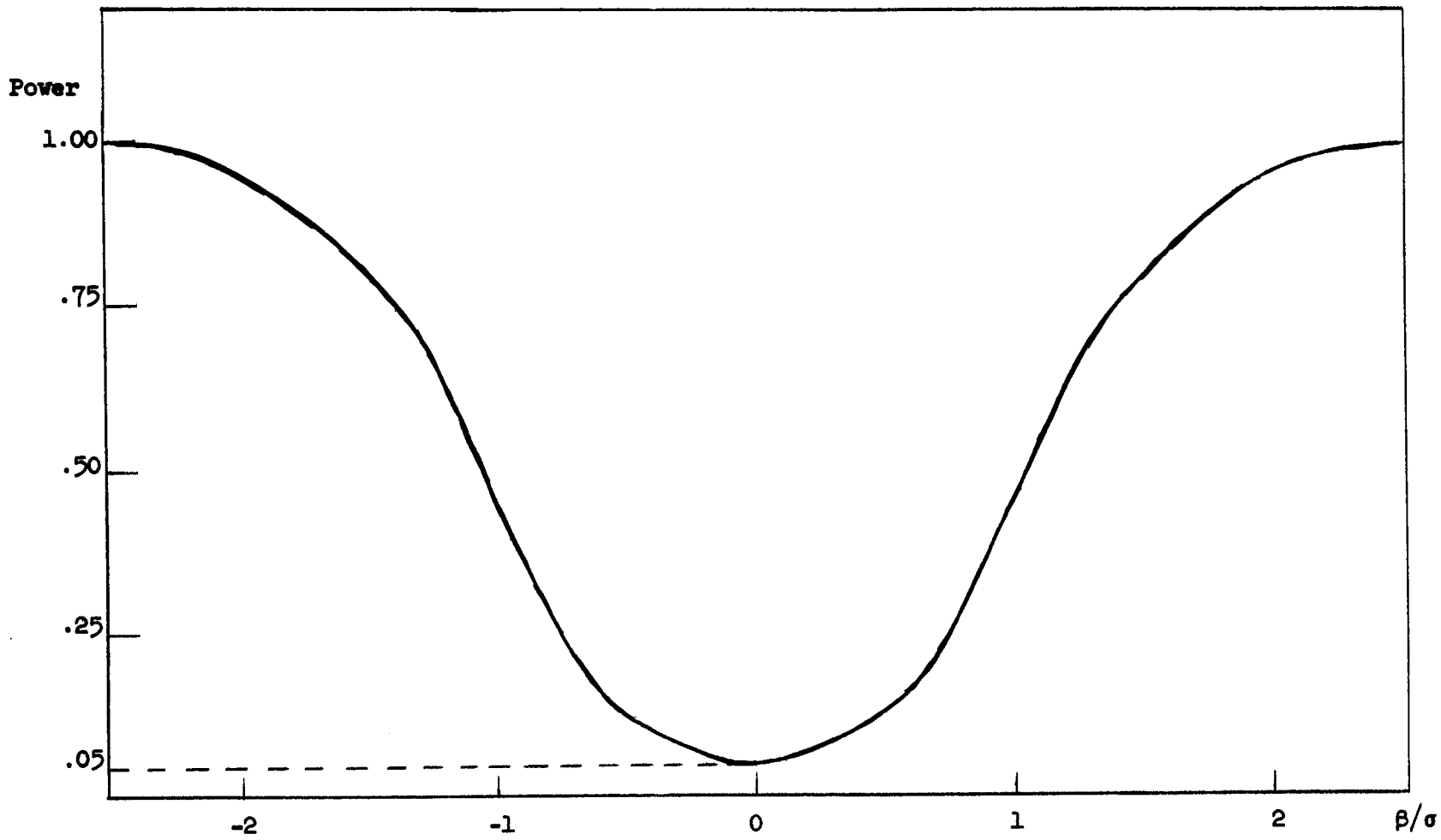

Figure 8

Power Function of Test of Hypothesis that $\beta = 0$

CHAPTER III

A TWO SAMPLE PROCEDURE

General Description

Assume the normal, linear regression model postulated in Chapter I. Let the non-stochastic variable x assume a fixed set of values x_i $(i = 1, 2, \ldots, p)$ such that each unit of sampling consists of p observations on y, corresponding to each of the p fixed values of x. Define

$$\bar{x} = \sum_{i=1}^{p} x_i/p \quad \text{and} \quad S = \sum_{i=1}^{p} (x_i - \bar{x})^2 \quad .$$

Assign a positive integer, n_o, to be the number of sampling units to be selected in a first sample. Let z be some positive, real number.

Let y_{ij} be the j-th $(j = 1, 2, \ldots, n_o)$ replicate of the observation on y corresponding to x_i $(i = 1, 2, \ldots, p)$. According to the model, the mean and variance of y_{ij} are

$$E(y_{ij}) = \alpha + \beta(x_i - \bar{x}) \quad (i = 1,\ldots,p;\ j = 1,\ldots,n_o) \quad , \tag{3.1}$$

and

$$\mathrm{Var}(y_{ij}) = \sigma^2 \quad (i = 1,\ldots,p;\ j = 1,\ldots,n_o) \quad , \tag{3.2}$$

respectively.

Preliminary estimates of α and β based on the n_o sampling units are

$$a^* = \frac{1}{p\, n_o} \sum_{i=1}^{p} \sum_{j=1}^{n_o} y_{ij} \quad , \tag{3.3}$$

and

$$b^* = \frac{1}{n_o\, S} \sum_{i=1}^{p} \sum_{j=1}^{n_o} y_{ij}(x_i - \bar{x}) \quad , \tag{3.4}$$

respectively. The variance σ^2 is estimated by

$$s^2 = \frac{1}{p\, n_o - 2} \left(\sum_{i=1}^{p} \sum_{j=1}^{n_o} y_{ij}^2 - p\, n_o\, a^{*2} - n_o\, S\, b^{*2} \right) . \tag{3.5}$$

Define the random variable n to be

$$n = \max \left\{ \left[\frac{s^2}{z} \right] + 1 \, , \, n_o \right\} \quad , \tag{3.6}$$

where q denotes the <u>largest integer less</u> than q . Select $(n-n_o)$ additional sampling units, making observations y_{ij} $(i = 1, 2, \ldots, p;$ $j = n_o + 1, n_o + 2, \ldots, n)$. Final estimates of α and β are

$$a = \frac{1}{pn} \sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij} \quad , \tag{3.7}$$

and

$$(3.8) \qquad b = \frac{1}{nS} \sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij}(x_i - \bar{x}) \quad ,$$

respectively. Now define

$$(3.9) \qquad t' = \frac{b\sqrt{nS}}{s} = \frac{(b - \beta)\sqrt{nS}}{s} + \frac{\beta\sqrt{nS}}{s} \quad .$$

Many of the theorems that follow will be simplified by application of

<u>Lemma 3.1</u>. Let s^2 be such that $\frac{n' s^2}{\sigma^2}$ has the χ^2 distribution with n' degrees of freedom and n be a random variable defined by (3.6). Let ξ be a random variable whose conditional distribution, given s and n, is normal with zero mean and variance σ^2/n. Then $\tau = \xi\sqrt{n}/s$ has Student's distribution with n' degrees of freedom unconditionally.

<u>Proof</u>. If s and n are fixed, then the conditional distribution of τ is normal with zero mean and variance σ^2/s^2. However n and s are random variables related by (3.6), which implies that $n = n_o$ is equivalent to

$$(3.10) \qquad n_o z \geqslant s^2 \quad ,$$

and $n = m > n_o$, similarly, is equivalent to

$$(3.11) \qquad m z \geqslant s^2 > (m - 1)z \quad .$$

As assumed in the hypothesis, s^2 is such that

$$\frac{n' s^2}{\sigma^2} = \chi^2 \quad , \tag{3.12}$$

has the χ^2 distribution with n' degrees of freedom:

$$p_{n'}(\chi^2) = \frac{1}{2^{n'/2}\Gamma(n'/2)} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{(n'-2)}{2}} \quad . \tag{3.13}$$

Thus

$$\begin{aligned} \text{Prob}\left\{n = n_o\right\} &= \text{Prob}\left\{s^2 \leqslant n_o z\right\} \\ &= \text{Prob}\left\{\frac{n' s^2}{\sigma^2} \leqslant \frac{n' z n_o}{\sigma^2}\right\} = \int_0^{\frac{n' z n_o}{\sigma^2}} p_{n'}(\chi^2) d\chi^2 \quad . \end{aligned} \tag{3.14}$$

Similarly, for $m \geqslant n_o + 1$,

$$\text{Prob}\left\{n = m\right\} = \int_{\frac{n'(m-1)z}{\sigma^2}}^{\frac{n' m z}{\sigma^2}} p_{n'}(\chi^2) d\chi^2 \quad . \tag{3.15}$$

The distribution function of τ may be written in conditional form with either s or n the conditional variable. It will be convenient to choose the latter. Thus

$$\text{Prob}\left\{|\tau| \leqslant \tau_o\right\} = \sum_{m=n_o}^{\infty} \text{Prob}_{n=m}\left\{|\tau| \leqslant \tau_o\right\} \text{Prob}\left\{n = m\right\}$$

(3.16)

$$= \sum_{m=n_o}^{\infty} \text{Prob}\left\{|\tau| \leqslant \tau_o \,,\, n = m\right\} \quad .$$

It was shown that the conditional distribution of τ was normal with zero mean and variance σ^2/s^2 which, by (3.12), equals n'/χ^2. Writing the variance of τ in this form and using (3.14),

$$\text{Prob}\left\{|\tau| \leqslant \tau_o, n = n_o\right\} =$$

(3.17)

$$\frac{1}{\sqrt{2\pi}\; 2^{n'/2}\,\Gamma(n'/2)} \int_{-\tau_o}^{\tau_o} \int_0^{\frac{n'zn_o}{\sigma^2}} (\chi^2/n')^{1/2}\, e^{-\frac{\chi^2 u^2}{2n'}}$$

$$\cdot\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}}\, d\chi^2 du \quad ,$$

and for $m > n_o$, using (3.15)

$$\text{Prob}\left\{|\tau| \leqslant \tau_o \,,\, n = m\right\} =$$

(3.18)

$$\frac{1}{\sqrt{2\pi}\; 2^{n'/2}\,\Gamma(n'/2)} \int_{-\tau_o}^{\tau_o} \int_{\frac{n'(m-1)z}{\sigma^2}}^{\frac{n'mz}{\sigma^2}} (\chi^2/n')^{1/2}$$

$$\cdot\, e^{-\frac{\chi^2 u^2}{2n'}}\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}}\, d\chi^2\, du \quad .$$

Substituting (3.17) and (3.18) into (3.16) one has

$$\text{Prob}\left\{|\tau| \leqslant \tau_o\right\} = \frac{1}{\sqrt{2\pi\, n'}\; 2^{n'/2}\,\Gamma(n'/2)}$$

$$(3.19) \quad \cdot \left[\int_{-\tau_o}^{\tau_o} \int_0^{\frac{n'zn_o}{\sigma^2}} (\chi^2)^{\frac{n'-1}{2}} \exp\left\{-\frac{\chi^2}{2}\left(1+\frac{u^2}{n'}\right)\right\} d\,\chi^2 du \right.$$

$$\left. + \sum_{m=n_o+1}^{\infty} \int_{-\tau_o}^{\tau_o} \int_{\frac{n'(m-1)z}{\sigma^2}}^{\frac{n'mz}{\sigma^2}} (\chi^2)^{\frac{n'-1}{2}} \exp\left\{-\frac{\chi^2}{2}\left(1+\frac{u^2}{n'}\right)\right\} d\,\chi^2 du \right].$$

The infinite summation in (3.19) is a uniformly convergent series in u. Therefore the summation may be permuted with the integration with respect to u, giving

$$\text{Prob}\left\{|\tau| \leqslant \tau_o\right\} = \frac{1}{\sqrt{2\pi\, n'}\; 2^{n'/2}\,\Gamma(n'/2)}$$

$$(3.20) \quad \cdot \left[\int_{-\tau_o}^{\tau_o} \int_0^{\frac{n'zn_o}{\sigma^2}} (\chi^2)^{\frac{n'-1}{2}} \exp\left\{-\frac{\chi^2}{2}\left(1+\frac{u^2}{n'}\right)\right\} d\,\chi^2 \; du \; + \right.$$

$$\left. \int_{-\tau_o}^{\tau_o} \left\{ \sum_{m=n_o+1}^{\infty} \int_{\frac{n'(m-1)z}{\sigma^2}}^{\frac{n'mz}{\sigma^2}} (\chi^2)^{\frac{n'-1}{2}} \exp\left\{-\frac{\chi^2}{2}\left(1+\frac{u^2}{n'}\right)\right\} d\,\chi^2 \right\} du \right] =$$

$$= \frac{1}{\sqrt{2\pi\, n'}\; 2^{n'/2}\,\Gamma(n'/2)} \int_{-\tau_o}^{\tau_o} \int_0^{\infty} e^{-\frac{\chi^2}{2}(1+\frac{u^2}{n'})} (\chi^2)^{\frac{n'-1}{2}} d\chi^2 du$$

$$= \frac{1}{\sqrt{\pi n'}\,\Gamma(n'/2)} \int_{-\tau_o}^{\tau_o} \int_0^{\infty} e^{-\frac{\chi^2}{2}(1+\frac{u^2}{2})} (\chi^2/2)^{\frac{n'-1}{2}} d(\chi^2/2) du$$

(3.20)

$$= \frac{(\frac{n'+1}{2})}{\sqrt{\pi n'}\,\Gamma(n'/2)} \int_{-\tau_o}^{\tau_o} (1 + \frac{u^2}{n'})^{\frac{n'+1}{2}} du = \int_{-\tau_o}^{\tau_o} g_{n'}(u) du \quad .$$

Equation (3.20) is just the probability that a variable having Student's distribution with n' degrees of freedom has absolute value $\leqslant \tau_o$. Hence τ has Student's distribution with n' degrees of freedom. This concludes the proof of the lemma.

The foregoing lemma may be applied directly to prove

<u>Theorem 3.1</u>. Let s, n and b be random variables defined by (3.5), (3.6) and (3.8) respectively. Let β be the unknown parameter of (3.1) and

$$S = \sum_{i=1}^{p} (x_i - \bar{x})^2 \quad .$$

Then

$$u = \frac{(b - \beta)\sqrt{nS}}{s} \quad ,$$

has Student's distribution with $n' = p\,n_o - 2$ degrees of freedom, and independent of σ^2, α and β.

Proof. Given n and s, b is conditionally normally distributed with mean β and variance σ^2/nS. Hence the conditional distribution of $\xi = (b - \beta)\sqrt{S}$ is normal with mean zero and variance σ^2/n. Since $u = \xi\sqrt{n}/s$, n is defined by (3.6) and $n's^2/\sigma^2$ has the χ^2 distribution with n' degrees of freedom, by Lemma 3.1, u has Student's distribution with n' degrees of freedom independent of σ^2, α and β.

Theorem 3.2. Assume the linear normal regression model and let a, b, s^2, and n be defined by (3.7), (3.8), (3.5) and (3.6) respectively. In addition to the n sampling units let there be m single observations on y, say y_{oj} $(j = 1, 2, \ldots, m)$, corresponding to a single unknown value x_o of x. Let $M = (\frac{1}{m} + \frac{1}{pn})$ and define $\bar{y}_o = \sum_{j=1}^{m} y_{oj}/m$. Then

$$(3.21) \qquad w = \frac{\bar{y}_o - a - b(x_o - \bar{x})}{s\sqrt{M + \frac{(x_o - \bar{x})^2}{nS}}},$$

has Student's distribution with $n' = p\,n_o - 2$ degrees of freedom.

Proof. Let n and s be defined by (3.6) and (3.5) respectively. Given n and s, then

$$(3.22) \qquad \xi = \frac{\bar{y}_o - a - b(x_o - \bar{x})}{\sqrt{n}\sqrt{M + \frac{(x_o - \bar{x})^2}{nS}}},$$

is conditionally normally distributed with zero mean and variance σ^2/n. Since $w = \xi\sqrt{n}/s$, and $n' s^2/\sigma^2$ has the χ^2 distribution with n' degrees of freedom, by Lemma 3.1, w has Student's distribution with $n' = (p\, n_o - 2)$ degrees of freedom.

Confidence Intervals in the Two Sample Procedure

It is necessary to show that Theorem 2.1 has an analog for the two sample procedure. Let γ, $0 \leqslant \gamma \leqslant 1$, be some preassigned number, $\nu = 1-\gamma/2$ and define $t_\nu(n')$ to be the 100ν percentile of Student's distribution with n' degrees of freedom. The notation of this chapter will continue to be used and let

$$B = b^2 - \frac{s^2\, t_\nu^2(n')}{nS} \quad . \tag{3.23}$$

Let k be the sample value of $(\bar{y}_o - a)$ and define two functions $L_i'(O_n)$ $(i = 1, 2)$ of the observations on the n sampling units by

$$L_1'(O_n) = \begin{cases} \bar{x} + \dfrac{b\,k}{B} - \dfrac{s\, t_\nu(n')}{B}\sqrt{\dfrac{k^2}{nS} + M\,B}\ , & \text{if } B > 0\ , \\[2ex] \bar{x} + \dfrac{k}{2b} - \dfrac{b\,S\,M}{2k}\ , & \text{if } B = 0 \text{ and } k > 0\ , \\[2ex] -\infty & \text{otherwise,} \end{cases} \tag{3.24}$$

and

$$(3.25)\quad L_2'(0_n) = \begin{cases} \bar{x} + \dfrac{b\,k}{B} + \dfrac{s\,t_\nu(n')}{B}\sqrt{\dfrac{k^2}{nS} + M\,B}\,, & \text{if } B > 0\,, \\[2ex] \bar{x} + \dfrac{k}{2b} - \dfrac{b\,S\,M}{2k}\,, & \text{if } B = 0 \text{ and } k < 0\,, \\[2ex] +\infty & \text{otherwise}\,. \end{cases}$$

The desired analog of Theorem 2.1 is provided by

<u>Theorem 3.3</u>. Let $L_i'(0_n)$ $(i = 1, 2)$ be two functions defined by (3.24) and (3.25). Let γ be a preassigned number such that $0 \leqslant \gamma \leqslant 1$ and let

$$B = b^2 - \frac{s^2\,t_\nu^2(n')}{nS}\,.$$

Regardless of the true value of x_o,

$$(3.26)\quad \begin{aligned} &\text{Prob}\left\{L_1'(0_n) \leqslant x_o \leqslant L_2'(0_n);\ x_o\right\} = \\ &\qquad 1 - \gamma + \gamma\left[1 - \text{Prob}\left\{B > 0\right\}\right]. \end{aligned}$$

<u>Proof</u>. Let

$$w = \frac{\bar{y}_o - a - b(x_o - \bar{x})}{s\sqrt{M + \dfrac{(x_o - \bar{x})^2}{nS}}}\,.$$

By Theorem 3.2, w has Student's distribution with $(p\,n_o - 2) = n'$

degrees of freedom. One may prove Theorem 3.3 in identically the same manner as used in proving Theorem 2.1. One need only replace the symbols n, S, u and $L_i(0_n)$ $(i = 1, 2)$ of Theorem 2.1 with pn, nS, w and $L_i'(0_n)$ $(i = 1, 2)$ respectively. The proof then follows verbatim; after using Theorem 3.2 as above, no further revision in the proof is required because of the two sample procedure.

<u>The Choice of z</u>

It is seen that efficiency in the two sample procedure reduces to making $\text{Prob}\left\{B > 0\right\}$ large. As defined in Theorem 3.3

$$(3.27) \qquad 1 - \text{Prob}\left\{B > 0\right\} = \text{Prob}\left\{ |b| \leqslant \frac{s\, t_\nu(n')}{\sqrt{nS}} \right\} .$$

Now given n and s, $(b - \beta)\, \sigma^{-1} \sqrt{nS}$ is normally distributed with zero mean and unit variance. Thus

$$(3.28) \qquad \begin{aligned} &\text{Prob}_{n,s}\left\{ |b| \leqslant \frac{s\, t_\nu(n')}{\sqrt{nS}} \right\} \\ &= \text{Prob}_{n,s}\left\{ -\frac{s\, t_\nu(n')}{\sigma} - \frac{\beta\sqrt{nS}}{\sigma} \leqslant \frac{b-\beta}{\sigma}\sqrt{nS} \leqslant \right. \\ &\qquad \left. \frac{s\, t_\nu(n')}{\sigma} - \frac{\beta\sqrt{nS}}{\sigma} \right\} \equiv \text{Prob}_{n,s}\left\{ A_1 \leqslant \frac{b-\beta}{\sigma}\sqrt{nS} \leqslant A_2 \right\} \\ &= (2\pi)^{-1/2} \int_{A_1}^{A_2} e^{-\frac{x^2}{2}}\, dx . \end{aligned}$$

From (3.6) it is seen that $n \geqslant s^2/z$ or $1/\sqrt{z} \leqslant \sqrt{n}/s$. Let β be outside some zone of indifference about zero, say $|\beta| \geqslant \ell$. Define

$C_i' = (-1)^i \dfrac{s\, t_\nu(n')}{\sigma} - \dfrac{\ell s \sqrt{S}}{\sigma \sqrt{z}}$ and $C_i = \dfrac{\sigma C_i'}{s}$ $(i = 1, 2)$. Since the normal probability density function is unimodal and symmetric about its mean,

$$(3.29)\qquad (2\pi)^{-1/2} \int_{A_1}^{A_2} e^{-\frac{x^2}{2}}\, dx \leqslant (2\pi)^{-1/2} \int_{C_1'}^{C_2'} e^{-\frac{x^2}{2}}\, dx =$$

$$\mathrm{Prob}_{n,s}\left\{ C_1' \leqslant \frac{(b-\beta)}{\sigma}\sqrt{nS} \leqslant C_2' \right\} = \mathrm{Prob}_{n,s}\left\{ C_1 \leqslant \frac{(b-\beta)}{s}\sqrt{nS} \leqslant C_2 \right\} .$$

Upon summing (3.25) over n and s as was done in the proof of Lemma 3.1, and since $\frac{(b-\beta)}{s}\sqrt{nS}$ has Student's distribution by Theorem 3.1, then

$$(3.30)\qquad 1 - \mathrm{Prob}\left\{ B > 0 \right\} \leqslant \int_{C_1}^{C_2} g_{n'}(u)\, du \quad .$$

The numbers ℓ, S and $t_\nu(n')$ are all constants. It is important to note that the parameter σ^2 does not enter into (3.30) either explicitly or implicitly. Equation (3.30) indicates the criterion on which to base the choice of z. To make $\mathrm{Prob}\left\{ B > 0 \right\} \geqslant 1 - \delta/\gamma$ whenever $|\beta| > \ell$, it suffices to fix z so that

$$(3.31)\qquad \int_{C_1}^{C_2} g_{n'}(u)du \leqslant \delta/\gamma < 1 .$$

An explicit value for z may be obtained by equating the upper limits to $t_{\delta/\gamma}(n')$ obtaining $C_2 = t_\nu(n') - \frac{\ell\sqrt{S}}{\sqrt{z}} = t_{\delta/\gamma}(n')$, so that

$$(3.32)\qquad z = \frac{\ell^2 S}{\left[t_\nu(n') - t_{\delta/\gamma}(n') \right]^2} \quad .$$

The resulting lower bound will generally not be distinguishable from $-\infty$ for computational purposes.

Exact determination of z, considering both upper and lower limits of the integral (3.31), may be obtained by using a table of the Student's distribution to approximate the equality, but for practical computation, the formulation in (3.32) is sufficient.

The Choice of n_o

The purpose of this section is to examine the criteria on which to base a choice of the initial sample size, n_o. It was seen in (3.14) that

$$\text{(3.33)} \qquad \text{Prob}\left\{n = n_o\right\} = \int_0^{\frac{n'zn_o}{\sigma^2}} p_{n'}(\chi^2)d\chi^2 ,$$

and in (3.15) that, for $m > n_o$,

$$\text{(3.34)} \qquad \text{Prob}\left\{n = m\right\} \int_{\frac{n'z(m-1)}{\sigma^2}}^{\frac{n'zm}{\sigma^2}} p_{n'}(\chi^2)d\chi^2 .$$

The expected value of n is obtained from

$$E(n) = \sum_{m=n_o}^{\infty} m \,\text{Prob}\left\{n = m\right\}$$

$$\text{(3.35)} \qquad = n_o \int_0^{\frac{n'zn_o}{\sigma^2}} p_{n'}(\chi^2)d\chi^2 + \sum_{m=n_o+1}^{\infty} \int_{\frac{n'z(m-1)}{\sigma^2}}^{\frac{n'zm}{\sigma^2}} m\, p_{n'}(\chi^2)d\chi^2 .$$

Let $\lambda = z/\sigma^2$ so that λ is a function of n' and hence of n_o. A typical term in the summation of (3.35) may be written as

$$(3.36) \qquad \frac{1}{2^{n'/2}\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} m\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}} d\chi^2 \quad .$$

For any $m > n_o$, the variable of integration, χ^2, is constrained by the limits so that

$$n'\lambda(m-1) \leqslant \chi^2 \leqslant n'\lambda m \quad ,$$

or

$$(3.37) \qquad \frac{\chi^2}{n'\lambda} \leqslant m \leqslant \frac{\chi^2}{n'\lambda} + 1 \quad .$$

Substituting (3.37) into (3.36), there results the inequality

$$\begin{aligned} & \frac{1}{2^{n'/2}\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} (\chi^2/n'\lambda)\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{(n'-2)}{2}} d\chi^2 \\ (3.38) \qquad & \leqslant \frac{1}{2^{n'/2}\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} m\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}} d\chi^2 \\ & \leqslant \frac{1}{2^{n'/2}\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} \left(\frac{\chi^2}{n'\lambda} + 1\right) e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}} d\chi^2 \quad . \end{aligned}$$

Observing that

$$2^{n'/2}\left[\Gamma(n'/2)\right] n' = 2^{n'/2}\left[\Gamma(n'/2)\right] 2\,\frac{n'}{2} =$$

(3.39)

$$2^{(n'+2)/2}\,\Gamma\left[(n'+2)/2\right] ,$$

one may write (3.38) as

$$\frac{\lambda^{-1}}{2^{(n'+2)/2}\,\Gamma((n'+2)/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} e^{-\frac{\chi^2}{2}} (\chi^2)^{n'/2}\, d\chi^2$$

$$\leqslant \frac{1}{2^{n'/2}\,\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}}\, d\chi^2$$

(3.40)

$$\leqslant \frac{\lambda^{-1}}{2^{(n'+2)/2}\,\Gamma((n'+2)/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} e^{-\frac{\chi^2}{2}} (\chi^2)^{n'/2}\, d\chi^2$$

$$+ \frac{1}{2^{n'/2}\,\Gamma(n'/2)} \int_{n'\lambda(m-1)}^{n'\lambda m} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n'-2}{2}}\, d\chi^2 ,$$

which is the same as

$$\frac{1}{\lambda} \int_{n'\lambda(m-1)}^{n'\lambda m} p_{n'+2}(\chi^2)\, d\chi^2 \leqslant \int_{n'\lambda(m-1)}^{n'\lambda m} m\, p_{n'}(\chi^2)\, d\chi^2$$

(3.41)

$$\leqslant \frac{1}{\lambda} \int_{n'\lambda(m-1)}^{n'\lambda m} p_{n'+2}(\chi^2)\, d\chi^2 + \int_{n'\lambda(m-1)}^{n'\lambda m} p_{n'}(\chi^2)\, d\chi^2 .$$

Substituting (3.41) into (3.35) and summing, one obtains

$$n_o \int_0^{n'\lambda n_o} p_{n'}(\chi^2)d\chi^2 + \frac{1}{\lambda} \int_{n'\lambda n_o}^{\infty} p_{n'+2}(\chi^2)d\chi^2$$

$$(3.42) \qquad \leqslant E(n) \leqslant n_o \int_0^{n'\lambda n_o} p_{n'}(\chi^2)d\chi^2 +$$

$$\frac{1}{\lambda} \int_{n'\lambda n_o}^{\infty} p_{n'+2}(\chi^2)d\chi^2 + \int_{n'\lambda n_o}^{\infty} p_{n'}(\chi^2)d\chi^2 ,$$

as bounds for $E(n)$. The bounds are good ones, for the difference between the upper and lower bound is just the last term which is less than one. It is seen that $E(n)$ is a function of n_o and λ and therefore of n_o , σ^2 , ℓ , γ and δ . If, to find that n_o which minimizes $E(n)$, one differentiated (3.42) with respect to n_o and equated the derivative to zero, there would result a formidable equation to be solved. It is much more convenient to use (3.42) to plot $E(n)$ against n_o for various choices of p and λ .

Table 1 contains $E(n)$ and the 95 per cent points $(n_{.95})$ of n as a function of n_o for the combinations of $p = 4$ and 8 and $\lambda = .1$ and .2 . As typical examples, the cases of $p = 8$ and $\lambda = .1$ and .2 are plotted in Figure 9. The expected values were calculated from the dominant inequality of (3.42) and the upper percentage points directly from (3.31) and (3.34). It may be seen that doubling p effects only a slight decrease in $E(n)$, but $n_{.95}$ is significantly

TABLE I

EXPECTED VALUE OF n , $E(n)$, AND THE .95 PER CENT POINT, $n_{.95}$, AS A FUNCTION OF n_o FOR VARIOUS COMBINATIONS OF p AND λ

n_o	$p=4$ $\lambda=.1$		$p=8$ $\lambda=.1$		$p=4$ $\lambda=.2$		$p=8$ $\lambda=.2$	
	$E(n)$	$n_{.95}$	$E(n)$	$n_{.95}$	$E(n)$	$n_{.95}$	$E(n)$	$n_{.95}$
1	10.99	29.95	11.00	20.98	5.95	14.98	5.98	10.49
2	10.99	20.98	11.00	16.91	5.95	10.49	5.98	8.46
3	10.99	18.31	11.00	15.45	5.95	9.16	5.96	7.73
4	10.99	16.91	11.00	14.60	5.97	8.46	5.91	7.30
5	10.99	16.04	11.00	14.04	6.11	8.02	5.92	7.02
6	10.99	15.45	10.99	13.60	6.49	7.73	6.28	6.80
7	10.99	14.90	10.98	13.41	7.16	7.45	7.04	7.00
8	11.04	14.60	10.97	13.20	8.03	8.00	8.00	8.00
9	11.15	14.32	10.98	12.90	9.00	9.00	9.00	9.00
10	11.38	14.04	11.12	12.73				
11	11.78	13.80	11.50	12.60				
12	12.39	13.60	12.16	12.48				
13	13.16	13.55	13.05	13.00				
14	14.05	14.00	14.01	14.00				
15	15.02	15.00	15.00	15.00				

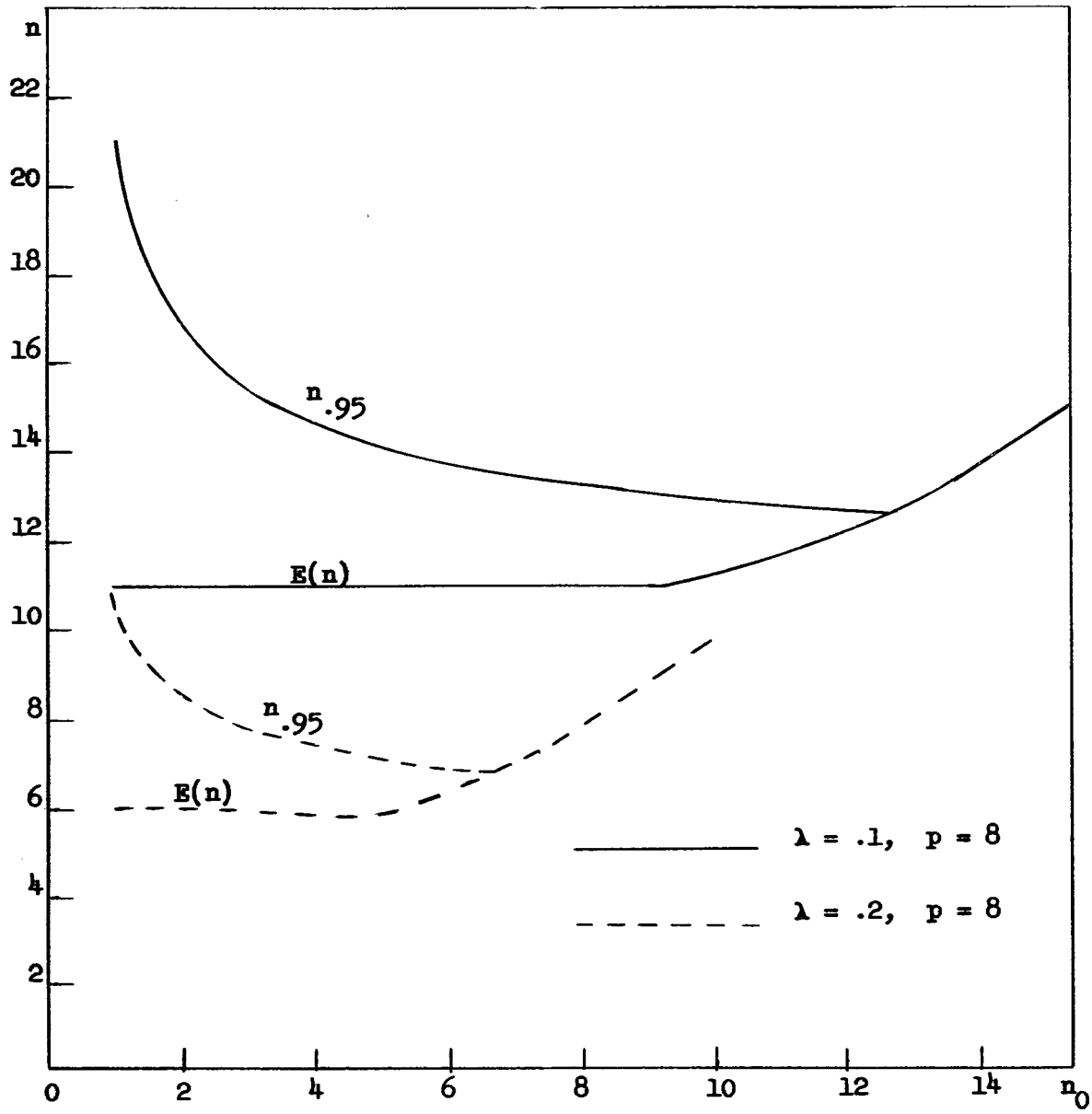

Figure 9

Expected Value, E(n), and 95 Per Cent Point, $n_{.95}$, of Distribution of n as a Function of n_0

reduced. On the other hand, doubling λ results in almost a 50 per cent reduction in $E(n)$ with exactly a 50 per cent decrease in $n_{.95}$ for the lower values of n_o .

The slight minima of the $E(n)$ curves are not detectable on the graph. It should be observed that when $p = 8$ and $\lambda = .2$ if n_o were chosen to be 4 , the expected total number of observations is 5.91 whereas if n_o were chosen to be 6 , $E(n) = 6.28$, an increase of $.37$ observations. However, in the first case, 5 per cent of the time in the long run, the total number of observations will exceed 7.30 whereas if $n_o = 6$, the upper 5 per cent point is only 6.80 , half an observation less. It may frequently be advantageous to choose n_o not as that value which minimizes $E(n)$, but so as to make $E(n)$ not greatly different from the absolute minimum and to provide thereby further insurance in the form of a smaller upper percentage point.

Choosing z and n_o is equivalent to designing the two sample procedure. If σ^2 is known, or an approximation exists, then z is chosen as a function of n_o , so that $\lambda = z/\sigma^2$ is a function of n_o . If $E(n)$ is too large for some practical reason, it is possible to adjust the specifications of the problem to reduce $E(n)$. The adjustments will consist of the alteration of γ , δ or ℓ or any combination of the three.

If σ^2 is not approximately known, so that λ is unknown, it is a more conservative procedure to choose n_o too large rather than too small. This safeguards one from the very high upper percentage point characteristic of the distribution of n for small n_o .

CHAPTER IV

MULTIPLE LINEAR NORMAL REGRESSION SINGLE SAMPLE

Introduction

This chapter will extend the usual one sample method of constructing a confidence interval for x_o to the case where there are more than one continuous, non-stochastic variables. In particular, the discussion will be based specifically on the model postulating two independent, arbitrary variables, but the formal extension to higher dimensions is straightforward.

Discrimination in Multiple, Linear, Normal Regression

Let y denote a normally distributed random variable with constant, unknown variance σ^2, and whose mean is an unknown linear function of two continuous, non-stochastic variables x_1 and x_2. Let there be n observations (y_j, x_{1j}, x_{2j}) $(j = 1, 2, \ldots, n)$ on y corresponding to n known pairs (x_{1j}, x_{2j}) of x_1 and x_2. Define $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$, $(i = 1, 2)$. Let the expected value of y be written in the form

$$E(y) = \alpha_o + \alpha_1(x_1 - \bar{x}_1) + \alpha_2(x_2 - \bar{x}_2) , \tag{4.1}$$

for arbitrary x_1 and x_2 and where α_o, α_1, and α_2 are unknown parameters. It will be convenient to introduce the notation

$$\ell_{ih} = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{hj} - \bar{x}_h) \quad (i, h = 1, 2) , \tag{4.2}$$

with the further provision that

$$(4.3) \qquad \ell_{oh} = \frac{1}{n} \sum_{j=1}^{n} y_j(x_{hj} - \bar{x}_h) \qquad (h = 1, 2) \quad ,$$

and

$$(4.4) \qquad \ell_{oo} = \frac{1}{n} \sum_{j=1}^{n} y_j \quad .$$

Let

$$(4.5) \qquad \Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \ell_{11} & \ell_{12} \\ 0 & \ell_{21} & \ell_{22} \end{bmatrix} \quad ,$$

and denote the determinant of Λ by L and the cofactor of ℓ_{ih} by L_{ih} $(i, h = 1, 2)$. By Cauchy's inequality [12; p. 2] $L \geqslant 0$ and it shall be assumed that $L > 0$. This is equivalent to the linear independence of x_1 and x_2 . Let a_i be an estimate of α_i $(i = 0, 1, 2)$ and define the column vectors

$$(4.6) \qquad a = \begin{bmatrix} a_o \\ a_1 \\ a_2 \end{bmatrix} \quad ,$$

and

$$(4.7) \qquad \ell_o = \begin{bmatrix} \ell_{oo} \\ \ell_{01} \\ \ell_{02} \end{bmatrix} .$$

Then a is the solution of the matrix equation

$$(4.8) \qquad \begin{aligned} \Lambda a &= \ell_o , \\ a &= \Lambda^{-1} \ell_o . \end{aligned}$$

Hence, the expectation of the a_i $(i = 0, 1, 2)$ is

$$E(a) = E(\Lambda^{-1} \ell_o) = \Lambda^{-1} E(\ell_o) ,$$

so that

$$(4.9) \qquad a - E(a) = \Lambda^{-1} \left[\ell_o - E(\ell_o) \right] .$$

The variance-covariance matrix of the estimates is

$$(4.10) \qquad \begin{aligned} & E\left\{ \left[a - E(a)\right] \left[a - E(a)\right]^T \right\} \\ &= E\left\{ (\Lambda^{-1}) \left[\ell_o - E(\ell_o)\right] \left[\ell_o - E(\ell_o)\right]^T (\Lambda^{-1})^T \right\} = \end{aligned}$$

$$(4.10)\quad \begin{aligned} &= (\Lambda^{-1})\, E\left\{\left[\ell_o - E(\ell_o)\right]\left[\ell_o - E(\ell_o)\right]^T\right\} (\Lambda^{-1})^T \\ &= \Lambda^{-1}\, \frac{\sigma^2}{n}\, \Lambda (\Lambda^{-1})^T = \frac{1}{n}\, \Lambda^{-1}\, \sigma^2 \quad , \end{aligned}$$

since $(\Lambda^{-1})^T = \Lambda^{-1}$ because of the symmetry of Λ. An estimate of σ^2 is

$$(4.11)\qquad s^2 = \frac{1}{n-3} \sum_{j=1}^{n} \left[y_j - a_o - a_1(x_{1j} - \bar{x}_1) - a_2(x_{2j} - \bar{x}_2)\right]^2 ,$$

which is independent of a_i $(i = 0, 1, 2)$ [1; p. 552].

A series of m observations y_{oh} $(h = 1, 2, \ldots, m)$ are taken, all corresponding to the unique, but unknown, pair (x_{10}, x_{20}). Let $\bar{y}_o = \sum_{h=1}^{m} y_{oh}/m$, and consider

$$(4.12)\qquad v = \bar{y}_o - a_o - a_1(x_{10} - \bar{x}_1) - a_2(x_{20} - \bar{x}_2) .$$

Since v is a linear combination of normally distributed variables, it too is normally distributed and, by virtue of (4.1) has mean zero. An estimate of the variance of v is

$$(4.13)\qquad u^2 = s^2 \left[\frac{1}{m} + \frac{1}{n} + \sum_{i,j=1}^{2} \frac{L_{ij}}{nL} (x_{io} - \bar{x}_i)(x_{jo} - \bar{x}_j)\right] ,$$

with $n' = n - 3$ degrees of freedom and independent of the parameter estimates in (4.12). Thus, the ratio $t = v/u$ has Student's distribution with n' degrees of freedom. If $M = (\frac{1}{m} + \frac{1}{n})$, $\nu = 1-\gamma/2$ and

$t_\nu(n')$ is the 100ν percentile of Student's distribution with n' degrees of freedom, then

$$(1 - \gamma) = \text{Prob}\left\{ |t| \leqslant t_\nu(n') \right\}$$

(4.14)

$$= \text{Prob}\left\{ \frac{\left| \bar{y}_o - a_o - a_1(x_{10} - \bar{x}_1) - a_2(x_{20} - \bar{x}_2) \right|}{s\sqrt{M + \sum_{i,j=1}^{2} \frac{L_{ij}}{nL}(x_{io} - \bar{x}_i)(x_{jo} - \bar{x}_j)}} \leqslant t_\nu(n') \right\}$$

Squaring and rearranging terms within the braces of (4.14), the latter may be written in the form

(4.15)

$$\text{Prob}\left\{ \sum_{i,j=1}^{2} \left[a_i a_j - \frac{s^2 t_\nu^2(n') L_{ij}}{nL}(x_{io}-\bar{x}_i)(x_{jo}-\bar{x}_j) \right] + (\bar{y}_o - a_o)^2 - 2\sum_{i=1}^{2} a_i(\bar{y}_o - a_o)(x_{io} - \bar{x}_i) - s^2 M t_\nu^2(n') \leqslant 0 \right\} = 1 - \gamma .$$

<u>A Geometric Construction</u>

The inequality within the braces of (4.15) may be taken to define a prediction region for $\bar{y}_o$ in terms of x_{io} $(i = 1, 2)$. To determine the nature of the bounding surface it is sufficient to evaluate the signature of the matrix of the quadratic form within the braces in (4.15). Let

$$(4.16) \qquad T_{ih} = \frac{s^2\, t_\nu^2(n')\, L_{ih}}{nL} \qquad (i,\ h = 1,\ 2)\ ,$$

then the matrix is

$$(4.17) \qquad \Delta = \begin{bmatrix} a_1^2 - T_{11} & -a_1 & a_1 a_2 + T_{12} \\ -a_1 & 1 & -a_2 \\ a_1 a_2 + T_{12} & -a_2 & a_2 - T_{22} \end{bmatrix} .$$

Let

$$(4.18) \qquad Q = \begin{bmatrix} 1 & \dfrac{a_1}{a_1^2 - T_{11}} & \dfrac{T_{12}}{T_{11}} \\ 0 & 1 & \dfrac{a_2 T_{11} + a_1 T_{12}}{T_{11}} \\ 0 & 0 & 1 \end{bmatrix} ,$$

so that

$$(4.19) \qquad Q^T \Delta Q = \begin{bmatrix} a_1^2 - T_{11} & 0 & 0 \\ 0 & -\dfrac{T_{11}}{a_1^2 - T_{11}} & 0 \\ 0 & 0 & \dfrac{T_{12}^2 - T_{11} T_{22}}{T_{11}} \end{bmatrix} .$$

The signature of $Q^T \Delta Q$ is -1 because $T_{ii} > 0$ $(i = 1, 2)$ and

$T_{12}^2 - T_{11}\, T_{22} = -s^4\, t_\nu^4(n')/n^2\, L < 0$. Since the signature is invariant under a congruence transformation, the signature of Δ is also -1 and inasmuch as $s^2\, M\, t_\nu^2(n') > 0$, the bounding surface is an hyperboloid of two sheets [3; p. 230] and the confidence region is that region of the space bounded by the two nappes of the hyperboloid.

A joint confidence region may be obtained for $(x_{10} - \bar{x}_1)$ and $(x_{20} - \bar{x}_2)$ corresponding to an observed value $\bar{y}_o - a_o = k$ by substitution into the expression within the braces of (4.15), obtaining

$$(a_1^2 - T_{11})(x_{10} - \bar{x}_1)^2 + 2(a_1\, a_2 + T_{12})(x_{10} - \bar{x}_1)(x_{20} - \bar{x}_2) +$$

$$(4.20) \qquad (a_2^2 - T_{22})(x_{20} - \bar{x}_2)^2 - 2\, a_1\, k(x_{10} - \bar{x}) - 2\, a_2\, k(x_{20} - \bar{x}_2)$$

$$+ k^2 - s^2\, t_\nu^2(n')\, M \leqslant 0 .$$

The equality in (4.20) represents a conic in the x_{10} , x_{20} plane whose discriminant is

$$D = 4\left[(a_1\, a_2 + T_{12})^2 - (a_1^2 - T_{11})(a_2^2 - T_{22})\right]$$

$$(4.21) \qquad = 4\left[a_1^2\, T_{22} + a_2^2\, T_{11} - T_{11}\, T_{22} + T_{12}^2 + 2\, a_1\, a_2\, T_{12}\right] .$$

To evaluate the sign of D one may consider the matrix Δ' formed by permuting the last two rows and columns of Δ . Thus

$$(4.22)\quad \Delta' = \begin{bmatrix} a_1^2 - T_{11} & a_1 a_2 + T_{12} & -a_1 \\ a_1 a_2 + T_{12} & a_2^2 - T_{22} & -a_2 \\ -a_1 & -a_2 & 1 \end{bmatrix} .$$

If

$$(4.23)\quad Q' = \begin{bmatrix} 1 & -\dfrac{a_1 a_2 + T_{12}}{a_1^2 - T_{11}} & \dfrac{a_1 T_{22} + a_2 T_{12}}{D/4} \\ 0 & 1 & \dfrac{a_2 T_{11} + a_1 T_{12}}{D/4} \\ 0 & 0 & 1 \end{bmatrix} ,$$

then

$$(4.24)\quad Q'^T \Delta' Q' = \begin{bmatrix} a_1^2 - T_{11} & 0 & 0 \\ 0 & -\dfrac{D/4}{(a_1^2 - T_{11})} & 0 \\ 0 & 0 & \dfrac{T_{12}^2 - T_{11} T_{22}}{D/4} \end{bmatrix} .$$

The signature of $Q'^T \Delta' Q'$ equals the signature of Δ' which must, in turn, be equal to that of Δ which is -1 . Since $T_{12}^2 - T_{11} T_{22} \leqslant 0$, then $\operatorname{sgn}(a_1^2 - T_{11}) = 1$ implies $\operatorname{sgn} D = 1$, but the converse does not hold.

The permutation of rows and columns in (4.22) is arbitrary for the quadratic form was originally symmetric in x_{10} and x_{20} . Thus

if either of $a_i^2 > T_{ii}$ for $i = 1$ or 2, or both, hold, $D > 0$ and the curve in the x_{10}, x_{20} plane is an hyperbola or one of its limiting forms, the result of the intersection of the plane $\bar{y}_o - a_o = k$ with both nappes of the hyperboloid. If both $a_1^2 > T_{11}$ and $a_2^2 > T_{22}$ the conditional joint confidence region for x_{10} and x_{20} will be taken to be the region in the plane bounded by the hyperbola (4.20). Since (4.20) is a direct consequence of (4.17), the associated confidence coefficient is $1 - \gamma$.

If either $a_1^2 > T_{11}$ or $a_2^2 > T_{22}$, but both inequalities do not hold simultaneously, a reduction to the simple regression model will be made. If, say, $a_2^2 \leqslant T_{22}$, then the conditional joint confidence region will be defined by letting $-\infty < x_{20} < \infty$ and by constructing an interval for x_{10} according to the discussion in Chapter II, identifying a_1, x_1 and x_{10} with b, x and x_o respectively.

If both $a_1^2 \leqslant T_{11}$ and $a_2^2 \leqslant T_{22}$, then no joint confidence region for x_{10} and x_{20} strictly exists. As in Case (2a) of Chapter II, one may <u>define</u> a conditional confidence region to be the entire x_{10}, x_{20} plane.

<u>The Analytic Determination of a Joint Confidence Region</u>

It will be demonstrated in this section that the heuristic discussion of the previous section may be used as a basis for a rigorous definition of an <u>unconditional</u> joint confidence region for (x_{10}, x_{20}). The discussion will refer to the following

<u>Definitions</u>. 1) <u>Condition P_1</u> will be said to obtain if

$a_1^2 > T_{11}$ and $a_2^2 > T_{22}$.

2) <u>Condition P_2</u> will be said to obtain if $a_i^2 > T_{ii}$ <u>and</u> $a_h^2 \leqslant T_{hh}$ $(i, h = 1, 2;\ i \neq h)$.

3) <u>Condition P_3</u> will be said to obtain if Conditions P_1 and P_2 do not, i.e., both $a_1^2 \leqslant T_{11}$ <u>and</u> $a_2^2 \leqslant T_{22}$ hold simultaneously.

4) If $\bar{y}_o - a_o = k$, then the <u>region</u> $\underline{R}$ is the set of all points (x_{10}, x_{20}) satisfying (4.20).

5) If $\bar{y}_o - a_o = k$, then a random covering region, dependent on the sample O_n , which is a subspace of the x_1 , x_2 space is defined to be

$$(4.25)\quad \Omega(O_n) = \begin{cases} \text{(i) all pairs } (x_{10}, x_{20}) \text{ in } R \text{ if Condition } P_1 \text{ obtains} \\ \text{(ii) } L_1(O_n) \leqslant x_{io} \leqslant L_2(O_n);\ -\infty < x_{ho} < \infty \text{ where } L_j(O_n)\ (j = 1, 2) \text{ are defined in (2.17) and (2.18) with } a_i,\ x_i \text{ and } x_{io} \text{ identified with } b,\ x \text{ and } x_o \text{ respectively if Condition } P_2 \text{ obtains} \\ \text{(iii) all pairs } (x_{10}, x_{20}) \text{ in the plane if Condition } P_3 \text{ obtains.} \end{cases}$$

The random covering region $\Omega(O_n)$ agrees with the limits set in the geometric discussion earlier in this chapter.

<u>Theorem 4.1</u>. Let Conditions P_1 , P_2 and P_3 be defined as in Definitions (1) - (3) above. Let $\Omega(O_n)$ be a random covering region defined by (4.25) depending on the sample O_n . Let γ be some

preassigned number such that $0 \leqslant \gamma \leqslant 1$. Then, regardless of the true values of x_{10} and x_{20},

$$\text{(4.26)} \quad \begin{aligned} &\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n); x_{10}, x_{20}\right\} \\ &= 1 - \gamma + \gamma\left[1 - \text{Prob}\left\{\text{Cond. } P_1 \text{ or Cond. } P_2\right\}\right] . \end{aligned}$$

<u>Proof</u>. Conditions P_1, P_2 and P_3 are exhaustive and mutually exclusive so that

$$\text{(4.27)} \quad \begin{aligned} &\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n); x_{10}, x_{20}\right\} \\ &= \text{Prob}\left\{(x_{10}) \in \Omega(0_n), \text{Cond. } P_1; x_{10}, x_{20}\right\} \\ &+ \text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n), \text{Cond. } P_2; x_{10}, x_{20}\right\} \\ &+ \text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n), \text{Cond. } P_3; x_{10}, x_{20}\right\} . \end{aligned}$$

Now

$$\text{(4.28)} \quad \begin{aligned} &\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n), \text{Cond. } P_1; x_{10}, x_{20}\right\} \\ &= \text{Prob}_{P_1}\left\{(x_{10}, x_{20}) \in \Omega(0_n); x_{10}, x_{20}\right\} \text{Prob}\left\{\text{Cond. } P_1\right\}. \end{aligned}$$

Let u and v be defined by (4.12) and (4.13) respectively. Then $t = u/v$ is distributed as $g_{n'}(t)$, Student's distribution with $n' = (n - 3)$ degrees of freedom. Let $\nu = 1 - \gamma/2$ and $t_\nu(n')$ the 100ν

percentile point of $g_{n'}(t)$. Then

$$1 - \gamma = \int_{-t_\nu(n')}^{t_\nu(n')} g_{n'}(t)dt = \text{Prob}\left\{t \leqslant t_\nu(n')\right\} - \text{Prob}\left\{t \leqslant -t_\nu(n')\right\} . \tag{4.29}$$

Now $(\bar{y}_o - a_o)$ is statistically independent of a_1 , a_2 and s , and t is clearly a monotone increasing function of $(\bar{y}_o - a_o)$. Thus, given $\bar{a} = (a_1, a_2, s)$ and letting

$$A_h = a_1(x_{10} - \bar{x}_1) + a_2(x_{20} - \bar{x}_2) + (-1)^h s\, t_\nu(n') \sqrt{M + \sum_{i,j=1}^{2} \frac{L_{ij}}{L} (x_{io} - \bar{x}_i)(x_{jo} - \bar{x}_j)} \quad (h = 1, 2) , \tag{4.30}$$

then, in view of the monotonicity of t ,

$$1 - \gamma = E_{\bar{a}}\left[\text{Prob}_{\bar{a}}\left\{\bar{y}_o - a_o \leqslant A_2; x_{10}, x_{20}\right\} - \text{Prob}_{\bar{a}}\left\{\bar{y}_o - a_o \leqslant A_1; x_{10}, x_{20}\right\}\right] =$$

$$E_{\bar{a}}\left[\text{Prob}_{\bar{a}}\left\{A_1 \leqslant \bar{y}_o - a_o \leqslant A_2; x_{10}, x_{20}\right\}\right] ,$$

independent of the true values of x_{10} and x_{20}. Let

$$\varphi(x_{io}, x_{ho}) = \frac{M}{\left[(x_{io} - \bar{x}_i) + \frac{L_{ih}}{L_{ii}}(x_{ho} - \bar{x}_h)\right]^2} + \frac{L_{ii}}{nL}$$

(4.32)

$$+ \frac{(L_{ii} L_{hh} - L_{ih}^2)\ (x_{ho} - \bar{x}_h)^2}{nL\ L_{ii}\left[(x_{io}-\bar{x}_i) + \frac{L_{ih}}{L_{ii}}(x_{ho}-\bar{x}_h)\right]^2} > \frac{L_{ii}}{nL} \quad ,$$

so that

$$(4.33) \qquad \frac{\partial A_j}{\partial(x_{io} - \bar{x}_i)} = a_i + (-1)^j \frac{s\ t_\nu(n')\ L_{ii}}{nL\sqrt{\varphi(x_{io}, x_{ho})}}$$

$$(i, h, j = 1, 2;\ i \neq h) \ .$$

If $a_i > 0$, then $\partial A_1/\partial(x_{io} - \bar{x}_i) > a_i - s\ t_\nu(n')\sqrt{L_{ii}/nL} > 0$ when $a_i^2 > T_{ii}$. Clearly $\partial A_2/\partial(x_{io} - \bar{x}_i) > 0$ if $a_i > 0$. If $a_i < 0$, then $\partial A_1/\partial(x_{io} - \bar{x}_i)$ is obviously negative and $\partial A_2/\partial(x_{io} - \bar{x}_i) < a_i + s\ t_\nu(n')\sqrt{L_{ii}/nL} < 0$ when $a_i^2 > T_{ii}$.

Hence, when Condition P_1 obtains, A_j $(j = 1, 2)$ are monotone functions of x_{10} and of x_{20} separately and thus are monotone functions of the two variables jointly. Therefore, corresponding to an observed $(\bar{y}_o - a_o) = k$, say, when $(\bar{y}_o - a_o)$ is bounded by A_1 and A_2 then $(x_{10}, x_{20}) \in \Omega(0_n)$ so that (4.31) may be

written in the form

$$(1 - \gamma) = E_{\bar{a}}\left[\text{Prob}_{\bar{a}}\left\{(x_{10}, x_{20}) \in \Omega(0_n);\ x_{10}, x_{20}\right\}\right]$$

(4.34)

$$= \text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n);\ x_{10}, x_{20}\right\} ,$$

independent of the true value of (x_{10}, x_{20}). But (4.34) was derived assuming that Condition P_1 obtains. Hence, from (4.28),

$$\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n),\ \text{Cond. } P_1;\ x_{10}, x_{20}\right\} =$$

(4.35)

$$(1 - \gamma)\ \text{Prob}\left\{\text{Cond. } P_1\right\} .$$

The second term on the right hand side of (4.27) may be evaluated, assuming $a_1^2 > T_{11}$ and $a_2^2 \leqslant T_{22}$, by writing

$$\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n),\ \text{Cond. } P_2;\ x_{10}, x_{20}\right\}$$

$$= \text{Prob}_{P_2}\left\{L_1(0_n) \leqslant x_{10} \leqslant L_2(0_n),\ -\infty < x_{20} < \infty;\ x_{10}, x_{20}\right\}\ \text{Prob}\left\{\text{Cond. } P_2\right\}$$

(4.36)

$$= \text{Prob}_{P_2}\left\{L_1(0_n) \leqslant x_{10} \leqslant L_2(0_n);\ x_{10}, x_{20}\right\} \cdot \text{Prob}\left\{\text{Cond. } P_2\right\} .$$

According to the definition in (4.25), a_1 , x_1 and x_{10} are identified with b , x and x_o of Chapter II respectively. Thus

$$b^2 = a_1^2 > T_{11} = \frac{s^2\, t_\nu^2(n')\, L_{11}}{nL} = \frac{s^2\, t_\nu^2(n')}{n\, L_{22} - \dfrac{L_{12}^2}{L_{11}}}$$

(4.37)

$$> \frac{s^2\, t_\nu^2(n')}{nL_{22}} = \frac{s^2\, t_\nu^2(n')}{\sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2} = \frac{s^2\, t_\nu^2(n')}{S} \quad .$$

Hence with the identification made, $a_1^2 > T_{11}$ implies that $B > 0$ so that, using the results of Theorem 2.1,

$$\text{Prob}_{P_2}\left\{L_1(0_n) \leqslant x_{10} \leqslant L_2(0_n);\ x_{10},\ x_{20}\right\} \text{Prob}\left\{\text{Cond. } P_2\right\}$$

(4.38) $$= \text{Prob}_{B>0}\left\{L_1(0_n) \leqslant x_{10} \leqslant L_2(0_n);\ x_{10},\ x_{20}\right\} \text{Prob}\left\{\text{Cond. } P_2\right\}$$

$$= (1 - \gamma)\ \text{Prob}\left\{\text{Cond. } P_2\right\} \quad .$$

In view of the symmetry between a_1 and a_2 , one obtains the same results if their roles were reversed.

The final term of (4.27) reduces readily to

(4.39) $$\text{Prob}\left\{(x_{10},\ x_{20}) \in \Omega(0_n),\ \text{Cond. } P_3;\ x_{10},\ x_{20}\right\} =$$

$$(4.39) \qquad = \text{Prob}\left\{\text{Cond. } P_3\right\} = (1 - \gamma + \gamma)\ \text{Prob}\left\{\text{Cond. } P_3\right\} ,$$

since if Condition P_3 obtains, $\Omega(0_n)$ is the entire (x_{10}, x_{20}) plane so that the joint event on the left hand side of (4.39) imposes no restriction on (x_{10}, x_{20}) and the equality results.

Finally, substituting (4.35), (4.38), and (4.39) in (4.27) one obtains

$$\begin{aligned}
\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega(0_n);\ x_{10}, x_{20}\right\} &= (1 - \gamma)\ \text{Prob}\left\{\text{Cond. } P_1\right\} \\
&+ (1 - \gamma)\ \text{Prob}\left\{\text{Cond. } P_2\right\} + (1 - \gamma + \gamma)\ \text{Prob}\left\{\text{Cond. } P_3\right\} \\
&= (1 - \gamma)\left[\text{Prob}\left\{\text{Cond. } P_1\right\} + \text{Prob}\left\{\text{Cond. } P_2\right\} + \right. \\
(4.40) \qquad & \qquad \left. \text{Prob}\left\{\text{Cond. } P_3\right\}\right] + \gamma\ \text{Prob}\left\{\text{Cond. } P_3\right\} \\
&= (1 - \gamma) + \gamma\left[1 - \text{Prob}\left\{\text{Cond. } P_1 \text{ or Cond. } P_2\right\}\right] ,
\end{aligned}$$

regardless of the true values of x_{10} and x_{20}. This concludes the proof.

A result has been obtained that is analogous to that reached in Chapter II. The term $\text{Prob}\left\{\text{Cond. } P_1 \text{ or Cond. } P_2\right\}$ is equivalent to the power of a test of the composite hypothesis that $\alpha_1 = \alpha_2 = 0$. Stein [11] has shown that no non-trivial one sample test of such an hypothesis exists whose power is independent of the variance.

The following chapter will outline a two sample procedure which will enable one to make the confidence coefficient arbitrarily close to $(1 - \gamma)$ whenever α_1 and α_2 are outside some region of indifference, a region whose center is at the origin.

CHAPTER V

MULTIPLE LINEAR NORMAL REGRESSION TWO SAMPLE

Introduction

A two sample procedure will be described for obtaining a confidence region for the independent variables in the case of discrimination in normal, multiple linear regression. In particular, it will be assumed that there exist two independent variables. The theory is formally extensible to more than two such variable in a straightforward manner. It will be shown that a confidence region for (x_{10}, x_{20}) may be constructed with an associated true confidence coefficient that is dependent on a function which is equivalent to the power of the test of the composite hypothesis $\alpha_1 = \alpha_2 = 0$. Finally, a method will be exhibited for making the actual confidence coefficient differ by an arbitrarily small positive quantity from some assigned positive number less than one whenever α_1 and α_2 are outside some zone of indifference.

The Non-Central F Distribution

Prior to the actual description of the two sample test and its properties, it will be convenient to introduce a distribution function that, in the multivariate case, plays the role that the non-central t distribution does in the univariate case. This function is the non-central F distribution function.

Let u_i $(i = 1, 2, \ldots, m)$ each have Student's distribution with n' degrees of freedom. Let c_i $(i = 1, 2, \ldots, m)$ be real

constants. Stein [11] shows that

$$F' = \sum_{i=1}^{m} \left[\frac{u_i}{\sqrt{n'}} + c_i \right]^2 , \tag{5.1}$$

has the non-central F distribution defined by

$$\text{Prob}\left\{F' \leqslant T\right\} = \Phi_{m,n'}(T, \sum_{i=1}^{m} c_i^2)$$

$$= \frac{\Gamma((m + n')/2)}{\sqrt{\pi}\,\Gamma(n'/2)\,\Gamma((m-1)/2)} \int_0^T \int_{-\sqrt{F'}}^{\sqrt{F'}} (F' - \rho^2)^{\frac{m-3}{2}} \tag{5.2}$$

$$\cdot \left(1 + F' + 2\rho\sqrt{\sum_{i=1}^{m} c_i^2} + \sum_{i=1}^{m} c_i^2\right)^{-\frac{m+n'}{2}} d\rho\, dF' .$$

If $\sum_{i=1}^{m} c_i^2 = 0$, the distribution reduces to that of the ratio of independent chi-square variables with m and n' degrees of freedom in numerator and denominator respectively.

General Description

Assume the multiple, normal, linear regression model postulated in Chapter IV. Let the non-stochastic variable x_i $(i = 1, 2)$ assume a fixed set of p values (x_{1j}, x_{2j}) $(j = 1, 2, \ldots, p)$ which shall constitute a sampling unit. All samples will be integral replicates of the sampling unit. Let y_{jk} be the k-th replicate of an observation on y corresponding to (x_{1j}, x_{2j}) . Define $\bar{x}_i = \sum_{j=1}^{p} x_{ij}/p$,

$(i = 1, 2)$. Assign a positive integer n_o to be the number of sampling units observed in a first sample and let z be some positive constant. Let

$$(5.3) \qquad \ell'_{ih} = \frac{1}{p} \sum_{j=1}^{p} (x_{ij} - \bar{x}_i)(x_{hj} - \bar{x}_h) \quad (i, h = 1, 2) \ ,$$

$$(5.4) \qquad \ell^*_{oh} = \frac{1}{n_o p} \sum_{j=1}^{p} \sum_{k=1}^{n_o} y_{jk}(x_{hj} - \bar{x}_h) \quad (h = 1, 2) \ ,$$

and

$$(5.5) \qquad \ell^*_{oo} = \frac{1}{n_o p} \sum_{j=1}^{p} \sum_{k=1}^{n_o} y_{jk} \ .$$

Let

$$(5.6) \qquad \Lambda' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \ell'_{11} & \ell'_{12} \\ 0 & \ell'_{21} & \ell'_{22} \end{bmatrix} \ ,$$

denote the determinant of Λ' by L' and the cofactor of ℓ'_{ih} by L'_{ih} $(i, h = 0, 1, 2)$. As in Chapter IV, it will be assumed that $L' > 0$. Let

$$(5.7) \qquad a^* = \begin{bmatrix} a_0^* \\ a_1^* \\ a_2^* \end{bmatrix} ,$$

and

$$(5.8) \qquad \ell_0^* = \begin{bmatrix} \ell_{00}^* \\ \ell_{01}^* \\ \ell_{02}^* \end{bmatrix} .$$

The a_i^* are preliminary estimates of α_i $(i = 0, 1, 2)$ and are the solutions of the matrix equation

$$(5.9) \qquad \Lambda' a^* = \ell_0^* ,$$

or

$$(5.10) \qquad a^* = \Lambda'^{-1} \ell_0^* .$$

An unbiased estimate, s^2, of the variance σ^2 is given by

$$s^2 = \frac{1}{p\,n_o - 3} \sum_{j=1}^{p} \sum_{k=1}^{n_o} \left[y_{jk} - a_o^* - a_1^*(x_{1j} - \bar{x}_1) - a_2^*(x_{2j} - \bar{x}_2) \right]^2$$

(5.11)

$$= \frac{p\,n_o}{p\,n_o - 3} \left[\frac{1}{p\,n_o} \sum_{j=1}^{p} \sum_{k=1}^{n_o} y_{jk}^2 - a_o^* \ell_{oo}^* - a_1^* \ell_{01}^* - a_2^* \ell_{02}^* \right] .$$

Let

(5.12) $$n = \max \left\{ \left[\frac{s^2}{z} \right] + 1 , n_o \right\} ,$$

where $[q]$ is the largest integer <u>less than</u> q. Select an additional $(n - n_o)$ sampling units. Define ℓ'_{ih} $(i, h = 0, 1, 2)$ analogously to ℓ^*_{ih}, replacing n_o by n in the denominator and in the upper limit of the summation. Then if a_i are final estimates of α_i $(i = 0, 1, 2)$ and

(5.13) $$a = \begin{pmatrix} a_o \\ a_1 \\ a_2 \end{pmatrix} ,$$

then a is the solution of the equation

$$\Lambda' a = \ell_o' \quad , \tag{5.14}$$

or

$$a = \Lambda'^{-1} \ell_o' \quad . \tag{5.15}$$

It follows that the expectation of a is

$$E(a) = E(\Lambda'^{-1} \ell_o') = \Lambda'^{-1} E(\ell_o') = \begin{bmatrix} \alpha_o \\ \alpha_1 \\ \alpha_2 \end{bmatrix} \quad . \tag{5.16}$$

<u>Theorem 5.1</u>. Let $a_i (i = 1, 2)$ be defined by (5.15) and s and n by (5.11) and (5.12) respectively. Then

$$t_i = \frac{(a_i - \alpha_i)\sqrt{np}}{s\sqrt{\frac{L'_{ii}}{L'}}} \qquad (i = 1, 2) \quad , \tag{5.17}$$

has Student's distribution with $n' = (p\, n_o - 3)$ degrees of freedom.

<u>Proof</u>. Expanding (5.15),

$$a_i = \frac{1}{n\,p\,L'} \left[L'_{ii} \sum_{j=1}^{p} \sum_{k=1}^{n} y_{jk}(x_{ij} - \bar{x}_i) + \right.$$

(5.18)

$$\left. L'_{ih} \sum_{j=1}^{p} \sum_{k=1}^{n} y_{jk}(x_{hj} - \bar{x}_h) \right] \qquad (i = 1,2;\ ih = 2) .$$

Hence given n , the conditional distribution of a_i $(i = 1, 2)$ is normal with mean α_i and variance

(5.19)
$$\mathrm{Var}_n(a_i) = \frac{L_{ii}}{n\,p\,L'}\,\sigma^2 .$$

Since the conditional distribution of

$$\xi_i = \frac{(a_i - \alpha_i)\sqrt{np}}{\sqrt{n}\sqrt{\frac{L_{ii}}{L'}}} ,$$

is normal with zero mean and variance σ^2/n and $\frac{n' s^2}{\sigma^2}$ has the χ^2 distribution with n' degrees of freedom, the theorem follows by Lemma 3.1 since

$$t_i = \frac{\xi_i \sqrt{n'}}{s} \qquad (i = 1, 2) .$$

The Construction of a Confidence Region

Theorem 5.2. Assume the normal, linear, multiple regression model and let s^2 , n , a_i $(i = 0, 1, 2)$ be defined by (5.11), (5.12) and (5.15) respectively. In addition to the n sampling units, of p

observations each, let there be m individual observations y_{oj} $(j = 1, 2, \ldots, m)$ on y corresponding to a unique, but unknown, pair (x_{10}, x_{20}) of (x_1, x_2). Let $\bar{y}_o = \sum_{j=1}^{m} y_{oj}/m$ and $M = \frac{1}{m} + \frac{1}{pn}$. Then

$$(5.20) \qquad t' = \frac{\bar{y}_o - a_o - \sum_{i=1}^{2} a_i(x_{io} - \bar{x}_i)}{s\sqrt{M + \sum_{i=1}^{2} \frac{L'_{ih}}{pnL'}(x_{io} - \bar{x}_i)(x_{ho} - \bar{x}_h)}}$$

has Student's distribution with $n' = (p\, n_o - 3)$ degrees of freedom.

<u>Proof</u>. Given n and s, the conditional distribution of

$$(5.21) \qquad \xi = \frac{\bar{y}_o - a_o - \sum_{i=1}^{2} a_i(x_{io} - \bar{x}_i)}{\sqrt{n}\sqrt{M + \sum_{i=1}^{2} \frac{L'_{ih}}{pnL'}(x_{io} - \bar{x}_i)(x_{ho} - \bar{x}_h)}}$$

is normal with zero mean and variance σ^2/n. But $n' s^2/\sigma^2$ is distributed as χ^2 with n' degrees of freedom. Hence, by Lemma 3.1, $t' = \frac{\xi\sqrt{n}}{s}$ has Student's distribution with n' degrees of freedom.

The next theorem to be proven is based on the following

<u>Definitions</u>.

1) $$T'_{ih} = \frac{s^2\, t_\nu^2(n')\, L'_{ih}}{pnL'} .$$

2) <u>Condition P'_1</u> is said to obtain if $a_1^2 > T'_{11}$ <u>and</u> $a_2^2 > T'_{22}$ hold simultaneously.

3) <u>Condition P_2'</u> is said to obtain if $a_i^2 > T_{ii}'$ and $a_h^2 \leqslant T_{hh}'$ $(i, h = 1, 2;\ i \neq h)$.

4) <u>Condition P_3'</u> is said to obtain if Conditions P_1' and P_2' do not.

5) If $\bar{y}_o - a_o = k$, the <u>region</u> <u>R'</u> of the x_1 , x_2 plane is the set of all pairs (x_{10}, x_{20}) such that

$$(a_1^2 - T_{11}')(x_{10} - \bar{x}_1)^2 + 2(a_1 a_2 + T_{12}')(x_{10} - \bar{x}_1)(x_{20} - \bar{x}_2) +$$

$$(a_2^2 - T_{22}')(x_{20} - \bar{x}_2)^2 - 2\, a_1\, k(x_{10} - \bar{x}_1) - 2\, a_2\, k(x_{20} - \bar{x}_2) +$$

(5.22)

$$k^2 - s^2\, M\, t_\nu^2(n') \leqslant 0 .$$

6) If $\bar{y}_o - a_o = k$, a random covering region of the x_1 , x_2 space, which is a function of the sample O_n , is defined to be

(5.23)
$$\Omega'(O_n) = \begin{cases} \text{(i) all pairs } (x_{10}, x_{20}) \text{ in } R' \text{ if Condition } P_1' \text{ obtains.} \\ \text{(ii) } L_1'(O_n) \leqslant x_{io} \leqslant L_2'(O_n);\ -\infty < x_{ho} < \infty \text{ where } L_j'(O_n)\ (j = 1, 2) \text{ are defined in (3.24) and (3.25) with } a_i, x_i \text{ and } x_{io} \text{ identified with } b, x \text{ and } x_o \text{ respectively if Condition } P_2' \text{ obtains.} \\ \text{(iii) all pairs } (x_{10}, x_{20}) \text{ in the plane if Condition } P_3' \text{ obtains.} \end{cases}$$

<u>Theorem 5.3</u>. Let γ be a preassigned number in the range $0 \leqslant \gamma \leqslant 1$. For any sample O_n, let a random covering region $\Omega'(O_n)$ be defined by Definition (6) and Conditions P_1', P_2' and P_3' by Definitions (2) - (4) respectively. Then, independent of the true value of (x_{10}, x_{20})

$$\text{Prob}\left\{(x_{10}, x_{20}) \in \Omega'(O_n);\ x_{10}, x_{20}\right\}$$

(5.24)

$$= (1 - \gamma) + \gamma\left[1 - \text{Prob}\left\{\text{Cond. } P_1' \text{ or Cond. } P_2'\right\}\right].$$

<u>Proof</u>. This theorem is the dual of Theorem 4.1. By Theorem 5.2, t' has Student's distribution with $(p\, n_o - 3)$ degrees of freedom. The region R' may be identified with the region defined by (4.20). The proof of Theorem 4.1 may be applied directly to this theorem if the primed quantities and functions are replaced by the analogous unprimed quantities and functions of Theorem 4.1 and pn of this theorem is identified with n of Theorem 4.1 and the recourse to Theorem 2.1 in the proof of Theorem 4.1 is replaced with recourse to Theorem 3.3.

Theorem 5.3 provides a method of constructing a confidence region for the joint estimation of x_{10} and x_{20}. The next section will show how the actual confidence coefficient may be made arbitrarily close to $(1 - \gamma)$, whenever α_1 and α_2 are outside some zone of indifference by the proper choice of the constant z.

The Choice of z and n_o

As was done in Chapter III, a criterion will be established by which z may be chosen in order that $\gamma\left[1 - \text{Prob}\left\{\text{Cond. } P_1' \text{ or Cond. } P_2'\right\}\right] \leqslant \delta < \gamma$ for $0 < \delta < \gamma$ when α_1 and α_2 are outside some zone of indifference.

Assume that n_o has, in some manner, been fixed. By Definitions (1) to (3)

$$1 - \text{Prob}\left\{\text{Cond. } P_1' \text{ or Cond. } P_2'\right\} = \text{Prob}\left\{a_1^2 < T_{11}' \text{ and } a_2^2 < T_{22}'\right\}$$

$$= \text{Prob}\left\{\frac{a_1^2\, pnL'}{s^2\, L_{11}'} < t_\nu^2(n') \text{ and } \frac{a_2^2\, pnL'}{s^2\, L_{22}'} < t_\nu^2(n')\right\}$$

$$\leqslant \text{Prob}\left\{\frac{\sum_{i=1}^{2}\left(\sqrt{\frac{pnL'}{L_{ii}'}}\, a_i\right)^2}{s^2} < 2\, t_\nu^2(n')\right\} \tag{5.25}$$

$$= \text{Prob}\left\{\sum_{i=1}^{2}\left[\frac{(a_i - \alpha_i)\sqrt{\frac{pnL'}{L_{ii}'}}}{s\sqrt{n'}} + \frac{\alpha_i\sqrt{\frac{pnL'}{L_{ii}'}}}{s\sqrt{n'}}\right]^2 < \frac{2\, t_\nu^2(n')}{n'}\right\} .$$

By Theorem 5.1,

$$\frac{(a_i - \alpha_i)}{s}\sqrt{\frac{pnL'}{L_{ii}'}} \qquad (i = 1, 2) ,$$

have Student's distribution with n' degrees of freedom and furthermore, $n' s^2/\sigma^2$ has the chi square distribution with n' degrees of freedom. Although

$$\frac{\alpha_i}{s\sqrt{n'}} \sqrt{\frac{pnL'}{L'_{ii}}} \qquad (i = 1, 2) \quad ,$$

are not constants, by (5.12),

$$(5.26) \qquad \left| \frac{\alpha_i}{s\sqrt{n'}} \sqrt{\frac{pnL'}{L'_{ii}}} \right| \geqslant \left| \frac{\alpha_i}{\sqrt{z}\sqrt{n'}} \sqrt{\frac{p\,L'}{L'_{ii}}} \right| \qquad (i = 1, 2) \quad .$$

Substituting (5.26) back into (5.25),

$$(5.27) \qquad \begin{aligned} 1 - \text{Prob}\left\{\text{Cond. } P'_1 \text{ or Cond. } P'_2\right\} \leqslant \text{Prob} \Bigg\{ \sum_{i=1}^{2} \Bigg[\frac{(a_i - \alpha_i)}{s\sqrt{n_o}} \frac{nL'}{L'_{ii}} \\ + \frac{\alpha_i}{\sqrt{z}\sqrt{n'}} \sqrt{\frac{pL'}{L'_{ii}}} \Bigg]^2 < \frac{2\, t_\nu^2(n')}{n'} \Bigg\} \quad . \end{aligned}$$

Comparison with (5.1) shows that the left hand side of the inequality within the braces on the right has the non-central F distribution with $m = 2$, $n' = p\, n_o - 3$ and

$$c_i = \frac{\alpha_i}{\sqrt{z}\sqrt{n'}} \sqrt{\frac{pL'}{L'_{ii}}} \qquad (i = 1, 2) \quad .$$

Let the zone of indifference, outside of which the confidence coefficient should be less than $(1 - \gamma + \delta)$, be the region bounded by

the ellipse

$$(5.28) \qquad \alpha_1^2 \left(\frac{pL'}{L'_{11}}\right) + \alpha_2^2 \left(\frac{pL'}{L'_{22}}\right) = r^2 \quad .$$

Points on the ellipse correspond to

$$\sum_{i=1}^{2} c_i^2 = \frac{r^2}{z\ n'} \quad .$$

In the notation of (5.2), z must be chosen so that

$$(5.29) \qquad \Phi_{2,n'} \left(\frac{2\ t_\nu^2(n')}{n'} \ , \ \frac{r^2}{z\ n'}\right) \leqslant \delta/\gamma \quad .$$

Given r^2 , γ , δ and $n' = p\ n_o - 3$, such a choice is always possible. Denote this choice by $z^*(n_o, \gamma, r^2, \delta)$. In principle, it is possible to tabulate z^* for various combinations of the parameters.

The distribution of n is the same for the multiple regression model as for the simple model. The choice of n_o should be based on the same criterion as discussed in Chapter III. The expected value of n is given by (3.40), but the value of z used in the determination of λ should be z^* rather than (3.32). If the minimum $E(n)$ is too large for all choices of n_o , it is possible to reduce $E(n)$ by a change in r^2 , γ , δ or some combination of these constants. Such a change is equivalent to an alteration in the specifications of the experimental design.

CHAPTER VI

SUMMARY AND EXAMPLE

Summary

Let y be a random variable, normally distributed with constant, unknown variance σ^2 and with a mean that is an unknown linear function $y = \alpha^* + \beta^* x$ of a non-stochastic, continuous variable x. The problem of discrimination assumes that there exist a set of observations on y corresponding to a single, unknown value, x_o, of x. It is desired to devise an experimental procedure which will enable one to construct a confidence region for x_o with confidence coefficient $1 - \gamma$, $0 \leqslant \gamma \leqslant 1$.

It is shown that the usual one sample procedure is inefficient. The confidence coefficient always exceeds $(1 - \gamma)$ by an amount that is a function of β^*/σ. A two sample experimental procedure is developed in Chapter III for constructing a confidence interval for x_o, independent of σ, which whenever β^* is outside some specified region of indifference about zero, will have a confidence coefficient not greater than $(1 - \gamma + \delta)$, where δ is an arbitrary positive number $< \gamma$. If $\beta^* = 0$, or is very small in absolute value, one should not try to discriminate x from y.

The two sample procedure enables one to make the confidence coefficient independent of σ at the expense of making the sample size a random variable dependent on σ. In the problem under consideration it is generally more desirable to permit the data to

fix the sample size in order to attain a given precision than to specify the sample size arbitrarily and then let the data fix the accuracy attained.

The discrimination problem in the case of normal, multiple, linear regression is discussed and found to be quite analogous to the simple regression model. The two sample procedure is extended specifically to the three dimensional case, but the formal extension to the n-dimensional situation, for arbitrary n, is immediate.

Example[2]

In the period following the cessation of hostilities in Europe during World War II, considerable concern was evinced in command circles about the awareness of the troops to the world situation and current events. The effectiveness of devoting regularly scheduled periods to orientation talks and discussions was questioned. Field trips to various commands were made to interview and examine samples of troops with regard to their knowledge of topical questions. The primary sampling unit was the battalion. From one infantry battalion a set of twelve test scores on a simple objective examination was received. The arithmetic mean of the test scores was 84. It was found that no reliance could be placed on the battalion information and education officer's estimate of orientation time in the unit.

---

[2]The data of this example were freely adapted by the writer from material in his possession while assigned to the Research Branch, Information and Education Division, Headquarters, United States Forces, European Theater, 1945-1946.

There were available four control infantry battalions, from another division, which had regularly scheduled orientation periods of 0, 2, 4, and 6 hours monthly. On the basis of a sample from the control battalions, a 95 per cent confidence interval for the unknown amount of orientation time in the test battalion was desired.

It was assumed that over the range of orientation time of practical interest, the distribution of test scores, y, was normal with a constant variance σ^2, and with a mean that was a linear function of orientation time, x, of the form (1.2). For practical purposes, this assumption was sufficiently valid.

It was undesirable to demand a larger sample from the control battalions than necessary to insure a confidence coefficient greater than .95. No knowledge of σ^2 was available. Furthermore, if the true regression coefficient was greater than two in magnitude, the confidence coefficient should not exceed .95 by more than .005.

It was arbitrarily decided to let n_o equal eight sampling units. Each unit consisted of one test score from each of the four control battalions. The data observed in the first sample appear in Table II.

In this example $p = 4$, $n_o = 8$, $n' = p\, n_o - 2 = 30$, and $S = 20$. The conditions of the problem specify that $\ell = 2$, $\gamma = .05$ and $\delta = .005$. This is sufficient information with which to choose z. With $\delta/\gamma = .1$, then $t_{.1}(30) = -1.310$ so that by (3.32),

$$z = \frac{(2)^2\, 20}{\left[2.042 + 1.310\right]^2} = 7.12 \quad .$$

TABLE II

TEST SCORES OF ENLISTED MEN IN FOUR INFANTRY BATTALIONS WITH DIFFERENT NUMBER OF ORIENTATION HOURS

Sampling Unit No.	0 Hours	2 Hours	4 Hours	6 Hours
1	65	74	70	89
2	61	66	69	93
3	79	70	89	100
4	67	71	65	82
5	58	56	87	95
6	49	61	75	85
7	72	68	82	90
8	58	70	93	93

Preliminary estimates of the parameters α and β are, from (3.3) and (3.4), $a^* = 75.063$ and $b^* = 4.675$ respectively. The variance σ^2 is estimated from (3.5) to be $s^2 = 67.019$. Thus, from (3.6),

$$n = \max \left\{ \left[\frac{67.019}{7.12} \right] + 1 , 8 \right\} = 10 .$$

Two additional sampling units are necessary. The additional data obtained appear in the supplement to Table II.

TABLE II (Supplement)

TEST SCORES OF ENLISTED MEN IN FOUR INFANTRY BATTALIONS WITH DIFFERENT NUMBER OF ORIENTATION HOURS

Sampling Unit No.	0 Hours	2 Hours	4 Hours	6 Hours
9	54	76	66	82
10	57	73	78	92

Final estimates of α and β are by (3.7) and (3.8) $a = 74.50$ and $b = 4.66$ respectively. Evaluation of (3.23) shows that $B = 20.32 > 0$ and hence the confidence interval for x_o , the unknown orientation time is provided by (3.26) using the first definitions of the functions $L_i'(0_n)$ $(i = 1, 2)$ in (3.24) and (3.25). The calculated confidence interval was $(4.24 \leqslant x_o \leqslant 6.11)$. The

true confidence coefficient depends on the unknown β . If the true value of β is 4.66 , then $C_1 = -9.852$ and $C_2 = -5.768$. The actual confidence coefficient, if $\beta = 4.66$ has an upper bound

$$.95 + .05 \int_{-9.852}^{-5.768} g_{30}(u)du < .95001 ,$$

which is well within the bound $.955$ specified in the problem.

BIBLIOGRAPHY

[1] Harald Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946.

[2] G. B. Dantzig, "On the non-existence of tests of "Student's" hypothesis having power functions independent of σ," Ann. Math. Stat., Vol. 11 (1940), pp. 186-192.

[3] Arnold Dresden, Solid Analytical Geometry and Determinants, John Wiley and Sons, 1952.

[4] Anders Hald, Statistical Theory with Engineering Applications, John Wiley and Sons, 1952.

[5] J. L. Hodges, Jr., "The selection of initial sample size in the Stein two sample procedure," unpublished dissertation, University of California, Berkeley, 1948.

[6] J. O. Irwin, "On the calculation of the error of biological assays," J. Hyg., Camb., Vol. 39 (1943), pp. 121-128.

[7] N. L. Johnson and B. L. Welch, "Applications of the non-central t-distribution," Biometrika, Vol. 31 (1940), pp. 362-389.

[8] A. M. Mood, Introduction to the Theory of Statistics, McGraw-Hill Book Co., 1950.

[9] Jerzy Neyman and E. S. Pearson, "Contributions to the theory of testing statistical hypotheses," Stat. Res. Memoirs, Vol. 1 (1936), pp. 1-37.

[10] Karl Pearson, Tables of the Incomplete Γ-Function, Cambridge University Press, 1946.

[11] Charles Stein, "A two-sample test for a linear hypothesis whose power is independent of the variance," Ann. Math. Stat., Vol. 16 (1945), pp. 243-258.

[12] Gabor Szegő, Orthogonal Polynomials, American Mathematical Society, 1939.

[13] S. S. Wilks, Mathematical Statistics, Princeton University Press, 1946.