ON CERTAIN METHODS OF ESTIMATING THE LINEAR STRUCTURAL RELATION[1]

By J. Neyman and Elizabeth L. Scott

University of California, Berkeley

1. Introduction and summary. The first part of this paper considers two methods of estimating the linear structural relation between two variables both of which are subject to "error"; the second part of the paper comments on a recently advanced procedure for constructing the confidence region for the slope of the structural relation.

In 1940 Wald [1] initiated a certain procedure for estimating the linear structural relation between two variables both of which are subject to "error." Wald's idea was extended by Nair and Banerjee [2] and later by Bartlett [3]. These procedures require some knowledge about the values of certain non-observable variables. When this knowledge is not available there is a temptation to substitute information derived from observations. One such method was considered by Wald who found sufficient conditions for the consistency of the resulting estimate. The purpose of the first part of the present paper is to find the necessary and sufficient conditions for two procedures with reference to a slightly more general case, namely, when the "errors" in the two observable variables may be correlated. The results obtained indicate that the two procedures, applied in the case of no additional knowledge about the values of the non-observable variables, will lead to consistent estimates of the slope of the structural relation in very exceptional cases only.

In 1949 Hemelrijk [4] described a novel procedure for constructing the confidence region of the slope of the linear structural relation in the case when the non-observable variables have unknown fixed values and the observations are made with "error" which has the same probability distribution at each point. The present paper considers this same procedure when there is no information about the fixed non-observable variables and also when these variables are random variables, and shows that the probability that the confidence region covers the true slope is the same as before but that the probability of covering any other slope is now the same as this probability of covering the true slope.

2. Statement of the first problem. Let ξ, u, and v denote random variables with $E(\xi)$ finite and with $E(u) = E(v) = 0$, and let ξ be independent of the pair u, v. In Method 2 below, assume also that ξ, u, and v have finite variances.

The three variables ξ, u, and v are assumed to be nonobservable. However, it is possible to observe the random variables x and y defined by

[1] Prepared with the partial support of the Office of Naval Research, and presented at the Chicago meeting of the Institute of Mathematical Statistics, December 29, 1950.

$$x = \xi + u,$$
$$(1) \qquad y = \alpha + \beta\xi + v,$$

where α and β are unknown constants. The relation $\eta = \alpha + \beta\xi$ is called the *linear structural relation* between the random variables x and y. The variables u and v are called the components of *error* although only part or, even, none of them need represent "error" in the strict interpretation.

We consider that n pairs of observations, say x_i, y_i, for $i = 1, 2, \cdots, n$, will be made on x and y. It will be assumed that the triplet (ξ_i, u_i, v_i) corresponding to the ith pair of observations is completely independent of all other such triplets. This will imply the independence of the pairs (x_i, y_i) and (x_j, y_j), $i \neq j$. For the first part of the paper (Sections 2–5) we consider that after the observations are obtained, they will be renumbered according to the magnitude of x so that $x_i \leqq x_{i+1}$ for $i = 1, 2, \cdots, n - 1$. However, this renumbering will not be assumed in Section 6.

Two different procedures for estimating β are considered.

METHOD 1. Fix two numbers $a \leqq b$ such that $P\{x \leqq a\} > 0$ and $P\{x > b\} > 0$. Let Z_1, W_1 denote the arithmetic mean of the x_i's and y_i's, respectively, for those pairs of observations for which $x_i \leqq a$, and Z_2, W_2 for those pairs for which $x_i > b$. As an estimate of β, consider, say, $b_1 = (W_2 - W_1)/(Z_2 - Z_1)$.

METHOD 2. Fix two proportions, $p_1 > 0$ and $p_2 > 0$, such that $p_1 + p_2 \leqq 1$ and then let $r = [np_1]$ and $s = [np_2]$. Denote by Z_3, W_3 the arithmetic mean of the x_i's and y_i's, respectively, for which $i = 1, 2, \cdots, r$; and by Z_4, W_4 the corresponding mean for $i = n - s + 1, n - s + 2, \cdots, n$. The estimate of β is then, say, $b_2 = (W_4 - W_3)/(Z_4 - Z_3)$.

Both of these methods are tempting in practical applications involving the estimation of the linear structural relation between two variables both of which are observed with "error". The purpose of the first part of the present paper is to investigate the necessary and sufficient conditions for the consistency of the estimate of β in these two procedures.

3. Necessary and sufficient conditions for the consistency of the estimate b_1.

We wish to compute the stochastic limit of the estimate b_1. In order to do this, we shall use a slight generalization of the well known theorem of Khintchine (see page 253 in [5]). The proof used for this lemma follows directly from an unpublished result of Robert F. Tate.

THEOREM 1. (GENERALIZED THEOREM OF KHINTCHINE.) *If $\{X_i\}$ is an infinite sequence of random variables, all independent and having the same distribution with $E(X_i) = \Xi$; further, if $\{\nu_n\}$ is an infinite sequence of integer-valued and positive random variables tending in probability to infinity (that is, such that, for any M, $\lim_{n\to\infty} P\{\nu_n < M\} = 0$), then as $n \to \infty$ the arithmetic mean of a random number ν_n of variables $X_1, X_2, \cdots, X_{\nu_n}$ converges in probability to Ξ,*

$$\lim_{n\to\infty} \text{p} \frac{1}{\nu_n} \sum_{i=1}^{\nu_n} X_i = \Xi.$$

Each of the two terms in the numerator of b_1 is a mean of a random number, say ν_n, of random variables all having the same distribution function with finite expected value. The same remark applies to the denominator of b_1. As $n \to \infty$, the variable ν_n tends in probability to infinity. It follows that, at the same time, each of the four means converges in probability to its expectation. Now, using the theorem of Slutsky (see page 255 in [5]), we see that the stochastic limit of b_1 is equal to $E(W_2 - W_1)/E(Z_2 - Z_1)$, provided this is finite. Thus, in the following we shall be concerned with the conditions under which $E(W_2 - W_1) = \beta E(Z_2 - Z_1)$.

We consider first the expected values,

$$E(Z_1) = E(x \mid x \leqq a) = E(\xi + u \mid \xi + u \leqq a),$$
$$E(Z_2) = E(x \mid x > b) = E(\xi + u \mid \xi + u > b).$$

In expression (1), we may set $\alpha = 0$ without loss of generality since we consider only differences, $W_2 - W_1$, etc. We then have

$$E(W_1) = E(y \mid x \leqq a) = \beta E(\xi \mid \xi + u \leqq a) + E(v \mid \xi + u \leqq a).$$

and similarly

$$E(W_2) = \beta E(\xi \mid \xi + u > b) + E(v \mid \xi + u > b).$$

Thus,

$$\begin{aligned} E(W_2 - W_1) \\ = \beta E(Z_2 - Z_1) - E(\beta u - v \mid \xi + u > b) + E(\beta u - v \mid \xi + u \leqq a). \end{aligned} \tag{2}$$

Let $f(u)$ denote the expected value of v given u fixed. Then the expectation of v may be rewritten in terms of $f(u)$. Thus, for example,

$$E(v \mid \xi + u > b) = E\{E[v \mid (\xi + u > b), u]\} = E[f(u) \mid \xi + u > b].$$

Now

$$\begin{aligned} E(W_2 - W_1) = \beta E(Z_2 - Z_1) \\ - E[\beta u - f(u) \mid \xi + u > b] + E[\beta u - f(u) \mid \xi + u \leqq a]. \end{aligned} \tag{3}$$

It is seen that the necessary and sufficient condition that b_1 be a consistent estimate of β is that, say,

$$I = E[\beta u - f(u) \mid \xi + u \leqq a] - E[\beta u - f(u) \mid \xi + u > b] = 0. \tag{4}$$

Since the value of β is unknown, it is of interest to ask for conditions which will preserve the consistency of the estimate of β no matter what the value of β, $-\infty < \beta < \infty$, may be. Let I_0 be the value of I when $\beta = \beta_0$ and I_1 be the value of I when $\beta = \beta_0 + 1$. Subtracting I_0 from I_1, we find that, say,

$$J = E(u \mid \xi + u \leqq a) - E(u \mid \xi + u > b) = 0 \tag{5}$$

is a necessary condition for the consistency of the estimate b_1 irrespective of the value of β. We shall see that $J = 0$ is also a sufficient condition.

Let $\Phi(\xi)$, $G(u)$, and $H(v)$ denote the distribution function of ξ, u, and v, respectively, and let $F(x)$ denote the resulting distribution function of x. Now we may write

$$\begin{aligned} E(u \mid \xi + u \leqq a) &= [F(a)]^{-1} \iint_{\xi + u \leqq a} u \, d\Phi(\xi) \, dG(u) \\ &= [F(a)]^{-1} \int_{-\infty}^{+\infty} u\Phi(a - u) \, dG(u) \\ &= [F(a)]^{-1} \int_{-\infty}^{+\infty} u[\Phi(a - u) - \Phi(a)] \, dG(u), \end{aligned}$$

since $E(u) = 0$. Similarly,

$$E(u \mid \xi + u > b) = [1 - F(b)]^{-1} \int_{-\infty}^{+\infty} u[\Phi(b) - \Phi(b - u)] \, dG(u).$$

Thus, in expression (5),

$$J = \int_{-\infty}^{+\infty} u \left[\frac{\Phi(a - u) - \Phi(a)}{F(a)} + \frac{\Phi(b - u) - \Phi(b)}{1 - F(b)} \right] dG(u).$$

It is easy to see that, unless both terms in the expression in square brackets are zero, the integrand is always negative. Thus, the necessary and sufficient conditions for $J = 0$ are

$$\begin{aligned} \Phi(a - u) - \Phi(a) &= 0, \\ \Phi(b - u) - \Phi(b) &= 0 \end{aligned} \tag{6}$$

for all values of u except for a set of probability zero.

Let (μ, ν) denote the shortest interval such that $P\{\mu \leqq u \leqq \nu\} = 1$. We know that $\mu \leqq 0 \leqq \nu$ since $E(u) = 0$. Then conditions (6) imply

$$\begin{aligned} \Phi(a - \nu) &= \Phi(a - \mu), \\ \Phi(b - \nu) &= \Phi(b - \mu); \end{aligned} \tag{7}$$

or

$$\begin{aligned} P\{a - \nu < \xi \leqq a - \mu\} &= 0, \\ P\{b - \nu < \xi \leqq b - \mu\} &= 0. \end{aligned} \tag{8}$$

Equations (8) are the necessary and sufficient conditions for $J = 0$ and, therefore, the necessary conditions for the consistency of the estimate b_1. We shall now prove that they are also sufficient. In order to do so, we consider

$$\beta J - I = E[f(u) \mid \xi + u \leqq a] - E[f(u) \mid \xi + u > b]$$

and show that when conditions (8) are satisfied then $\beta J - I = 0$. Under conditions (8) we have

$$\beta J - I = E[f(u) \mid \xi \leqq a - \nu] - E[f(u) \mid \xi > b - \mu],$$

and, since the pair of random variables u, v is independent of the random variable ξ and since $E(v) = 0$, we have $\beta J - I = 0$. Hence the conditions (8) are the necessary and sufficient conditions for $I = 0$ irrespective of the value of β. We now have proved the following theorem:

THEOREM 2. *In order that b_1 preserve the property of being a consistent estimate of β irrespective of the value of β, $-\infty < \beta < \infty$, it is necessary and sufficient that $P\{a - \nu < \xi \leqq a - \mu\} = P\{b - \nu < \xi \leqq b - \mu\} = 0$.*

4. Necessary and sufficient conditions for the consistency of the estimate b_2. We now compute the stochastic limit of the estimate b_2. Since each average in the expression for b_2 is taken over dependent observations, the theorem of Khintchine is not directly applicable. We shall evaluate the expectation and variance of each average, shall show that each variance tends to zero and thus that each average converges in probability to its expected value.

Letting x_j denote the jth of the observed x_i's, $i = 1, 2, \cdots, n$, numbered, as above, in order of magnitude, we have

$$E(x_j) = nC_{n-1}^{j-1} \int_{-\infty}^{+\infty} x[F(x)]^{j-1}[1 - F(x)]^{n-j}\, dF(x). \tag{9}$$

Then

$$\begin{aligned} E(Z_4) &= \frac{n}{s} \sum_{j=n-s+1}^{n} C_{n-1}^{j-1} \int_{-\infty}^{+\infty} x[F(x)]^{j-1}[1 - F(x)]^{n-j}\, dF(x) \\ &= \frac{n}{s} \int_{-\infty}^{+\infty} x I_{F(x)}(n - s, s)\, dF(x), \end{aligned} \tag{10}$$

where $I_{F(x)}(n - s, s)$ is the incomplete Beta function,

$$I_{F(x)}(n - s, s) = \frac{\int_0^{F(x)} t^{n-s-1}(1 - t)^{s-1}\, dt}{\int_0^1 t^{n-s-1}(1 - t)^{s-1}\, dt}.$$

Let X_{1-p_2} denote the $(1 - p_2)$-percentile point of the distribution of x. As is well known, when $n \to \infty$ with $s = [np_2]$ where p_2 is fixed, then $I_{F(x)}(n - s, s)$ tends to zero for all $x < X_{1-p_2}$ and to unity elsewhere. Thus

$$\lim_{n\to\infty} E(Z_4) = \frac{1}{p_2} \int_{X_{1-p_2}}^{+\infty} x\, dF(x) = E(x \mid x > X_{1-p_2}). \tag{11}$$

We now need to show that $\sigma_{Z_4}^2 \to 0$ as $n \to \infty$ with $s = [np_2]$ where p_2 is fixed. Consider

$$s^2 Z_4^2 = \sum_{j=n-s+1}^{n} x_j^2 + 2 \sum_{i=n-s+1}^{n-1} \sum_{j=i+1}^{n} x_i x_j. \tag{12}$$

Repeating the reasoning leading to (10), we find

$$E\left(\sum_{j=n-s+1}^{n} x_j^2\right) = n \int_{-\infty}^{+\infty} x^2 I_{F(x)}(n - s, s)\, dF(x).$$

Further, for $j > i$,

$$E(x_i x_j) = n(n - 1)C_{n-2}^{i-1} C_{n-i-1}^{j-i-1} \int_{-\infty}^{+\infty} x[F(x)]^{i-1} \cdot \int_{x}^{\infty} z[F(z) - F(x)]^{j-i-1} [1 - F(z)]^{n-j}\, dF(z)\, dF(x).$$

Thus,

$$2 \sum_{i=n-s+1}^{n-1} \sum_{j=i+1}^{n} E(x_i x_j)$$
$$= 2n(n - 1) \int_{-\infty}^{+\infty} x \int_{x}^{\infty} z \sum_{i=n-s+1}^{n-1} C_{n-2}^{i-1} [F(x)]^{i-1} [1 - F(x)]^{n-i-1}\, dF(z)\, dF(x)$$
$$= 2n(n - 1) \int_{-\infty}^{+\infty} x I_{F(x)}(n - s, s - 1) \int_{x}^{\infty} z\, dF(z)\, dF(x).$$

Substituting into (12), we have

$$E(Z_4^2) = \frac{n}{s^2} \int_{-\infty}^{+\infty} x^2 I_{F(x)}(n - s, s)\, df(x) + 2\frac{n(n - 1)}{s^2} \int_{-\infty}^{+\infty} x I_{F(x)}(n - s, s - 1) \int_{x}^{\infty} z\, dF(z)\, dF(x). \tag{13}$$

Letting $n \to \infty$ with $s = [np_2]$ where p_2 is fixed, we obtain

$$\lim_{n\to\infty} E(Z_4^2) = \frac{2}{p_2^2} \int_{X_{1-p_2}}^{\infty} x \int_{x}^{\infty} z\, dF(z)\, dF(x)$$
$$= \frac{1}{p_2^2} \int_{X_{1-p_2}}^{\infty} \int_{X_{1-p_2}}^{\infty} z\, dF(z)\, dF(x)$$
$$= [\lim_{n\to\infty} E(Z_4)]^2.$$

Therefore,

$$\lim_{n\to\infty} \sigma_{Z_4}^2 = 0,$$

and it follows that as n is increased Z_4 converges in probability to $E(x \mid x > X_{1-p_2})$. Similarly, it can be shown that Z_3 converges in probability to $E(x \mid x \leqq X_{p_1})$.

We now want to compute $E(W_4)$. Let y_j be the observation accompanying x_j, the jth of the x's in order of magnitude. Then

$$E(y_j) = E[E\,(y_j \mid x_j)],$$

and we have

$$E(y_j) = \int_{-\infty}^{+\infty} E(y \mid x)\, nC_{n-1}^{j-1} [F(x)]^{j-1} [1 - F(x)]^{n-j}\, dF(x),$$

so that

$$E(W_4) = \frac{n}{s} \int_{-\infty}^{+\infty} E(y \mid x) \sum_{j=n-s+1}^{n} C_{n-1}^{j-1} [F(x)]^{j-1} [1 - F(x)]^{n-j} dF(x)$$

$$= \frac{n}{s} \int_{-\infty}^{+\infty} E(y \mid x)\, I_{F(x)}(n - s, s)\, dF(x).$$

Owing to the property of $I_{F(x)}(n - s, s)$ already mentioned, we thus have that, as $n \to \infty$ with $s = [np_2]$ where p_2 is fixed,

$$\lim_{n\to\infty} E(W_4) = \frac{1}{p_2} \int_{X_{1-p_2}}^{\infty} E(y \mid x)\, dF(x) = E(y \mid x > X_{1-p_2}).$$

Combining the reasoning above with that used to obtain $E(Z_4^2)$, it is easy to show that

$$\lim_{n\to\infty} \sigma_{W_4}^2 = 0.$$

It follows that, as n is increased, W_4 converges in probability to $E(y \mid x > X_{1-p_2})$. In a similar way, we can show that W_3 converges in probability to $E(y \mid x \leqq X_{p_1})$.

Now, noticing that the stochastic limits of W_3, W_4, Z_3, and Z_4 are identical with the expectations of W_1, W_2, Z_1, and Z_2, respectively, we can use the results obtained in Section 3 to establish Theorem 3.

Let $r = [np_1]$, $s = [np_2]$ and let ξ_{p_1} and ξ_{1-p_2} be the corresponding percentile points of ξ, that is, such that $P\{\xi \leqq \xi_{p_1}\} = p_1$ and $P\{\xi > \xi_{1-p_2}\} = p_2$.

THEOREM 3. *If $n \to \infty$ while p_1 and p_2 are held constant, the necessary and sufficient condition that b_2 preserve the property of being a consistent estimate of β irrespective of the value of β, $-\infty < \beta < \infty$, is that*

$$P\{\xi_{p_1} - \nu < \xi \leqq \xi_{p_1} - \mu\} = P\{\xi_{1-p_2} - \nu < \xi \leqq \xi_{1-p_2} - \mu\} = 0.$$

5. Remarks. A. Wald [1] considered estimates similar to b_2 for the case u and v uncorrelated, with $p_1 = p_2 = \frac{1}{2}$, and showed that the conditions in Theorem 3 are sufficient conditions.

It is interesting that b_1 and b_2 may be consistent estimates of β for *some* values of this parameter even though the conditions of the theorems are not satisfied. To see this, we return to formula (4). If u and v are dependent random variables and if the regression of v on u is represented by the equation $f(u) = \beta^* u$, then b_1 and b_2 are consistent estimates whenever $\beta = \beta^*$. On the other hand, if u and v are independent, then the terms in (4) involving $f(u)$ drop out because $E(v) = 0$, and we find that whenever $\beta \neq 0$ the necessary and sufficient conditions for the consistency of the estimates b_1 and b_2 are just the conditions in Theorems 2 and 3, respectively. However, if $\beta = 0$ then b_1 and b_2 are certainly consistent estimates of β.

The results obtained suggest that the two procedures discussed will lead to consistent estimates of β in very exceptional cases only.

6. On Hemelrijk's confidence region for β. Hemelrijk [4] considers the following construction of a confidence region, say B, for the slope β of the linear structural relation. Let (x_r, y_r) and (x_s, y_s) be two different points chosen from the n observable points in any manner which is completely independent of the u_i, v_i for $i = 1, 2, \cdots, n$. Consider the set B of values of the slope of two parallel straight lines, one through (x_r, y_r) and the other through (x_s, y_s), such that inside of the closed strip bounded by these two lines there are less than $n - m$ observed points. Hemelrijk shows, under the conditions stated at the beginning of Section 2, with the additional assumptions that (a) whatever the fixed numbers θ and p, the probability is zero that the errors u and v will satisfy the relation $u \cos \theta + v \sin \theta = p$ and (b) the ξ_i, for $i = 1, 2, \cdots, n$, are unknown fixed numbers, that the probability that the set B includes β is given by

$$P\{\beta \in \varepsilon, B\} = 1 - \frac{(m+1)(m+2)}{n(n-1)} \qquad \text{for } 0 \leqq m \leqq n-3. \tag{15}$$

It should be emphasized that unless some additional information, not assumed here, is available about the unobservable variables then, in order to fulfill the condition that the choice of the two points used to construct B be made in a manner which is completely independent of the (u_i, v_i) for $i = 1, 2, \cdots, n$, this choice ordinarily will be made at random out of the n observed points. Also in many practical situations it does not seem appropriate to consider the values of ξ as fixed constants. Rather, they are treated as independent samples of a random variable.

We now show that in either of these two cases, (i) when the values of ξ are fixed constants but the choice of the observed points to be designated r and s is made in a random manner, and (ii) when ξ is a random variable, the probability that the set B includes any fixed slope, say γ, is exactly the same as the probability that B includes the true slope β, as given by (15). The theorem is stated for the second case but the proof is identical in the two cases and is the same as that used by Hemelrijk to prove (15).

THEOREM 4. *Whatever the fixed number γ (whether coinciding with the slope β of the structural relation or not) under the conditions given at the beginning of Section 2 plus the condition (a) above, the probability that the set B includes γ is*

$$P\{\gamma \varepsilon B\} = 1 - \frac{(m+1)(m+2)}{n(n-1)} \qquad \textit{with } 0 \leqq m \leqq n-3. \tag{16}$$

PROOF. The set B includes γ if and only if the parallel lines of slope γ through the two points (x_r, y_r) and (x_s, y_s) determine a closed strip which contains fewer than $n - m$ points (x_i, y_i). Let z_i denote the distance, in an arbitrary fixed direction different from γ, from (x_i, y_i) to any fixed line L of slope γ. $\gamma \varepsilon B$ if and only if fewer than $n - m$ of the $z_1, z_2, \cdots, z_n$ lie in the closed interval $[z_r, z_s]$. Under the assumptions made, $z_1, z_2, \cdots, z_n$ are independent observations of the same random variable z (under the conditions of Hemelrijk, this is true only when γ coincides with β), with probability one that the z_i's are all different. Thus, the probability that z_r is the jth smallest of the z_i's is the same $1/n$ for every j. The same is true for z_s. The z_i's may be arranged in $n!$ ways. The number of arrangements for which fewer than $n - m$ of the z_i's lie in the closed interval $[z_r, z_s]$ is $n! - 2[(m + 1)(n - 2)! + m(n - 2)! + \cdots + (n - 2)!]$ so that the desired probability is

$$P\{\gamma \varepsilon B\} = 1 - \frac{(m + 1)(m + 2)}{n(n - 1)} \qquad \text{with } 0 \leqq m \leqq n - 3.$$

The theorem just proved implies that, under the conditions stated, the power of the test of the hypothesis that $\beta = \beta_0$, say, provided by the set B, is a constant equal to the probability of an error of the first kind.

The authors wish to emphasize that it is not their intention to criticize the elegant construction of Hemelrijk, which is perfectly correct in relation to the hypotheses he makes. The point under discussion is that, just as in the case of the result of Wald, one may feel tempted to apply Hemelrijk's procedure somewhat beyond the limits indicated. The results obtained here show that such extensions are not profitable. Since the distinction between the conditions assumed by Hemelrijk and those at the outset of this paper is somewhat delicate, some illustrations may be interesting.

(i) Consider that N astronomers propose to study the slope β of the structural relation between two characteristics, ξ and $\eta = \alpha + \beta\xi$, of the stars. Each astronomer will observe, independently from the others, the same n stars and will use his set of n pairs of observations, $\{x_i, y_i\}$, $i = 1, 2, \cdots, n$, to construct the confidence set B for the slope β. Furthermore, for the construction of B each will designate the observations from the same two stars chosen from the n stars in advance, say Castor and Pollux, as (x_r, y_r) and (x_s, y_s), respectively, and will use the same value for m. We have here the conditions assumed by Hemelrijk. Expression (15) holds but not necessarily expression (16) for $\gamma \neq \beta$. Thus, the expected proportion of the N sets B which include the true slope β is $1 - (m + 1)(m + 2)[n(n - 1)]^{-1}$, but this is not true, in general, for any other slope $\gamma \neq \beta$.

(ii) Consider a situation similar to that described in (i) except that each of the N astronomers chooses for himself, and in a random manner, the particular two stars, out of the n stars, to be designated as the rth and the sth in the construction of his set B. Now, whatever the number γ, whether coinciding with the true slope β or not, the expected proportion of the sets B which include γ is the same number, given by (16).

The same conclusion holds if this situation (ii) is altered by having each astronomer choose for himself the particular n stars that he will observe. However, if we consider the subset of the N astronomers, each choosing for himself the n stars that he will observe, who use the *same* two stars, say Castor and

Pollux, to construct the set B, then (16) does not necessarily hold for $\gamma \neq \beta$; we have the same conclusion as with situation (i).

(iii) The situations above may be considered somewhat unrealistic. Presumably, the N astronomers would not each construct his own confidence set B for the same slope β. Rather, their observations would be combined and then one set B constructed from the combined observations. We may, however, consider the cases in general human experience in which, each for its own problem, a set B will be constructed. The expected proportion of these sets B which include any number γ is exactly the same as the expected proportion which include the true slope, as given by Theorem 4.

REFERENCES

[1] A. Wald, "The fitting of straight lines if both variables are subject to error," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 284–300.

[2] K. R. Nair and K. S. Banerjee, "A note on fitting of straight lines if both variables are subject to error," *Sankhyā*, Vol. 6 (1942), p. 331.

[3] M. S. Bartlett, "Fitting a straight line when both variables are subject to error," *Biometrics*, Vol. 5 (1949), pp. 207–212.

[4] J. Hemelrijk, "Construction of a confidence region for a line," *Nederl. Akad. Wetensch., Proc.*, Vol. 52 (1949), pp. 374–384.

[5] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.