NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.
Confidence Bands in Straight Line Regression

A. V. Gafarian

17 April 1963
Confidence Bands in Straight Line Regression

by

A. V. Gafarian

17 April 1963

SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA
Confidence Bands in Straight Line Regression
by
A. V. Gafarian

ABSTRACT

This paper develops a method for obtaining confidence bands in polynomial regression when the observations are independently distributed with constant but unknown variance. The bands may be obtained, in principle, over arbitrary sets of the independent variable with exact preassigned confidence coefficients. In general, difficult distribution problems result when specific applications are attempted. The major portion of this paper is concerned with first degree polynomials since some progress has been made here. A table is provided to obtain a constant width confidence band which contains the true but unknown straight regression line for values of the independent variable in some arbitrarily selected interval with an exact preassigned confidence coefficient. The present method is compared with the classical hyperbolic band for the whole regression line.

1The author wishes to express his indebtedness to Mr. Vance A. Griffitts who did all the programming for the table.
1. **INTRODUCTION AND SUMMARY**

The basic problem considered in this paper is the following. Suppose for every \( t \in (-\infty, \infty) \), \( Y_t \) is a normal random variable with unknown variance \( \sigma^2 \) and mean value \( m_t \) given by a polynomial of known degree \( r \geq 1 \) and unknown coefficients. Let \( I \) be a subset of interest in \( (-\infty, \infty) \). Based on mutually independent observations it is desired to construct simultaneous confidence intervals for \( m_t \), \( t \in I \), with preassigned probability \( 1 - \alpha \). It should be pointed out that the material discussed here is close to methods called "multiple comparisons" in other contexts.

A well known result occurs when the set \( I \) contains only one point, Graybill [1, pp. 121-122]. It must be emphasized that if intervals are computed by that technique for every \( t \), no confidence statement may be made about the resulting band (a hyperbola for \( r=1 \)) containing the unknown regression line, i.e., that method does not provide simultaneous coverage of the ordinates of the regression line. Less known is the work of Working and Hotelling [2] in which a hyperbolic confidence band is obtained for the whole regression line when it is assumed the variance is known. The method is easily extended to the unknown variance case and provides a hyperbolic band valid for the whole regression line, Scheffé [3, pp. 52,53]. Hoel [4] extends the method of Working and Hotelling for the straight line regression in such a way as to make it possible to find an optimum confidence band. The optimum band is defined to be that band of an admissible class of bands such that its expected total area is a minimum. Also, in [4] the case of polynomial regression of degree two or higher is considered and a procedure similar to the first degree case is outlined. However, in these cases the confidence bands possess confidence coefficients \( \geq 1 - \alpha \).
The present study was undertaken to extend some of the results described above. Ordinarily an experimenter is not interested in coverage of the whole regression curve. On the contrary, interest lies in only a bounded interval or even a finite set of points. The restriction of the above described bands to bounded sets of interest yield confidence coefficients \( \geq 1 - \alpha \) (even in the first degree case). A method for providing a band that is valid only for the set of interest may yield a more efficient band. Secondly, it would be desirable to maintain a uniform degree of accuracy over the set of interest, i.e., the width of the band is the same for all values of the independent variable \( t \) in the set of interest.

This paper develops a general method for obtaining confidence bands of arbitrary shape and over any arbitrary subset of the line when the observations are independently normally distributed. The shape is arbitrary in the sense that if \( w \) is any positive function defined over the subset \( I \) of interest in \( (-\infty, \infty) \), then the width of the band for \( t \in I \) is proportional to \( w(t) \). Thus, by selecting \( w(t) = 1, t \in I \), the resulting band has the same width for every \( t \in I \).

In general, difficult distribution problems result when specific applications are attempted. The major portion of this paper is concerned with first degree polynomials since some progress has been made here. A table is provided to obtain a confidence band which contains the true regression line for values of the independent variable in an arbitrarily selected interval of interest \( [a, b] \) with an exact confidence coefficient. The band has the same width for all values \( t \in [a, b] \). The table is constructed for use in the following situation: (1) The sample size \( n \) is even; (2) If observations are made at the values
\[ t_1, t_2, \ldots, t_n \text{ of the independent variable then } \bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i = \frac{a+b}{2}. \]

Defining \([A,B]\) as the interval in which observations are permissible a best solution obtains if in addition (3) \(\frac{A+B}{2} = \frac{a+b}{2}\), i.e., the observation interval \([A,B]\) is symmetrically located with respect to the interval of interest \([a,b]\); (4) (2) is realized by making half the observations at \(A\) and half at \(B\). The solution is best in the sense that for a given \(n\), \(\frac{(B-A)}{(b-a)}\), and probability of coverage this particular experimental configuration achieves the smallest bandwidth.

The important feature of the band provided by the present method is that it is uniformly wide over \([a,b]\). In order to get some idea of its efficiency it was compared to the band that arises by merely considering the restriction of the hyperbolic one to the interval \([a,b]\), though in this case the probability of coverage is no longer \(1-\alpha\) but \(> 1 - \alpha\). The comparison was made in terms of the areas of the bands. To be more specific for a given \(n\), \(\frac{(B-A)}{(b-a)}\), and probability of coverage, the best band (i.e., minimum area) was computed by the present method. The experimental configuration to achieve this also provides the minimum area over \([a,b]\) for the hyperbolic band. The ratio of the two areas was then considered as a measure of the efficiency. Roughly, the result is that for \(\frac{(B-A)}{(b-a)} > \frac{3}{2}\) the present method is more efficient and for \(\frac{(B-A)}{(b-a)} < \frac{3}{2}\) the restriction of the hyperbolic band to \([a,b]\) yields smaller areas. More specific calculations will be presented in a later section of the paper.
2. GENERAL TECHNIQUE

Suppose that for every $t \in (-\infty, \infty)$, $Y_t$ is a normal random variable with unknown variance $\sigma^2$ and expectation given by a polynomial $\beta_0 + \beta_1 t + \ldots + \beta_r t^r$ of unknown coefficients and known degree $r$. Let $I \subset (-\infty, \infty)$ be the set of interest. For preassigned confidence coefficient $1-\alpha$ and positive function $w$ defined on $I$ it is desired to obtain simultaneous confidence intervals for $E[Y_t] = m_t$, $t \in I$, such that the length of the interval for each $t \in I$ is proportional to $w(t)$.

Suppose independent observations are made at the time points $t_1, t_2, \ldots, t_n$ where the number of distinct observation points is $\geq r + 1$ and the number of observations is $> r + 1$ (this ensures that $\sigma^2$ may be estimated since only $r + 1$ distinct points are needed for the estimability of the linear parameters). Let $\hat{\beta}' = (\hat{\beta}_0 \hat{\beta}_1 \ldots \hat{\beta}_r)$ denote the vector of least squares estimates for $\beta' = (\beta_0 \beta_1 \ldots \beta_r)$ given by

$$\hat{\beta} = (T'T)^{-1}T'Y$$

where

$$T = \begin{pmatrix}
1 & t_1 & t_1^2 & \ldots & t_1^r \\
1 & t_2 & t_2^2 & \ldots & t_2^r \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & t_n & t_n^2 & \ldots & t_n^r
\end{pmatrix}$$

and $Y' = (y_1, y_2, \ldots, y_n)$ is the vector of observations at the points $t_1, t_2, \ldots, t_n$. 
Denote by

$$\sigma^2 = \frac{1}{n-r-1} (Y-T\hat{\hat{\sigma}})'(Y-T\hat{\hat{\sigma}})$$

the independent unbiased estimator of $\sigma^2$ based on $n-r-1$ degrees of freedom. Let $\hat{\theta}_t$ be the best linear estimate of $m_t$ given by $\sum_{j=0}^{t} \hat{\theta}_j j$. From the function

$$\frac{\hat{\theta}_t - m_t}{w(t)\hat{\hat{\sigma}}}$$

and for any pair of numbers $(\delta_1, \delta_2)$ with $\delta_1 < \delta_2$ let

$$V(\delta_1, \delta_2) = \left\{ \frac{\hat{\theta}_t - m_t}{w(t)\hat{\hat{\sigma}}} : \delta_1 < \frac{\hat{\theta}_t - m_t}{w(t)\hat{\hat{\sigma}}} < \delta_2, \ t \in I \right\}$$

in the space of the random vector $\hat{\theta}_t - m_t$, whose distribution is parameter free and calculable [5]. These are sufficient conditions to obtain

$$[\hat{\theta}_t - \delta_2 w(t)\hat{\hat{\sigma}}, \hat{\theta}_t - \delta_1 w(t)\hat{\hat{\sigma}}], \ t \in I,$$

as simultaneous confidence intervals of confidence coefficient $P[V(\delta_1, \delta_2)]$ [6]. The width of the band for any $t \in I$ is $(\delta_2 - \delta_1) w(t)\hat{\hat{\sigma}}$.

To insure the existence of at least one pair $(\delta_1, \delta_2)$ to acquire the probability $1-\alpha$, an additional restriction must be imposed on the function $w$. The set $V(\delta_1, \delta_2)$ may be written as

$$\bigcap_{t \in I} \left\{ \frac{\hat{\theta}_t - m_t}{w(t)\hat{\hat{\sigma}}} : \delta_1 < \frac{\hat{\theta}_t - m_t}{w(t)\hat{\hat{\sigma}}} < \delta_2 \right\}.$$
Since

$$\hat{\theta}_t - m_t = \frac{1}{w(t)} \left( 1 \ t^2 \ ... \ t^r \right) \left( \frac{\hat{\theta} - \theta}{\sigma} \right)'$$

it follows that each set in the above intersection consists of the points between two parallel hyperplanes which are perpendicular to \((1 \ t \ ... \ t^r)\)' and are at distances \((w(t) |\delta_2|)/(\sum_{j=0}^{r} |t|^{1/2})\) and \((w(t) |\delta_1|)/(\sum_{j=0}^{r} |t|^{1/2})\) from the origin.

Hence, if there exist constants \(m > 0\) and \(M > 0\) such that \(m < w(t)/(\sum_{j=0}^{r} |t|^{1/2}) < M\) for \(t \in I\) then and only then does there exist a pair \((\alpha_1, \alpha_2)\) (actually many pairs) such that the required probability is attained.

It is conjectured that optimum confidence intervals are obtained whenever \(\delta_2\) is taken \(> 0\) and \(\delta_1 = -\delta_2\). The optimum is in the sense that for a given confidence coefficient \(1-\alpha\) the difference \(\delta_2 - \delta_1\), and hence the length of the confidence intervals, will be minimized. This conjecture is based on: (1) The fact that the density function for the random vector \(\frac{\hat{\theta} - \theta}{\sigma}\) is constant on concentric \((r+1)\)-dimensional ellipsoids with center at origin and decreases monotonely with distance from the origin, and (2) The set \(V(\alpha_1, \alpha_2)\) in this situation is symmetrical with respect to the origin and probably has a maximum volume for any fixed difference \(\delta_2 - \delta_1\).

It should be emphasized again that the real difficulty here is the calculation of \(\delta_1\) and \(\delta_2\) to achieve probability \(1-\alpha\) when any specific applications are attempted. Progress has been made for the case \(r=1\), \(I\) an interval, and \(w(t) = 1\) for \(t \in I\), i.e., a band which has the same width over the interval of interest.

The major portion of the remainder of the paper is devoted to this problem.
However, for some special examples the general case specializes properly to well known results. E.g., if \( I \) is a single point only, say \( t_0 \), \( w(t_0) = 1 \), \( \delta_1 = -\delta_2 \), and \( \delta_2 > 0 \), it can be shown that

\[
\delta_2 = t_0 \frac{(1 t_0 \ldots t_0^r)(T'T)^{-1}(1 t_0 \ldots t_0^r)'}{2;n-r-1} \frac{1}{2},
\]

where \( t_0 \frac{\alpha}{2};n-r-1 \) is the upper \( \alpha/2 \) point of a \( t \)-variable with \( n-r-1 \) degrees of freedom, so that

\[
P[\frac{\sum t^j}{j=0^r} - \delta_2 \leq \frac{\sum t^j}{j=0^r} t_0 \leq \frac{\sum t^j}{j=0^r} t_0 + \delta_2] = 1-\alpha,
\]

[1, p. 122]. Similarly, consider the set of all linear combinations

\[
\{\beta_0 u_0 + \ldots + \beta_r u_r: (u_0 u_1 \ldots u_r) \in E_{r+1}\}. \text{ Setting } \delta_1 = -\delta_2 \text{ and } \delta_2 > 0 \text{ and defining } w \text{ for any } (u_0 u_1 \ldots u_r) \text{ to equal}
\]

\[
[(u_0 u_1 \ldots u_r)(T'T)^{-1}(u_0 u_1 \ldots u_r)']^{1/2}
\]
gives that

\[
\delta_2 = (r+1) F_{\alpha;r+1, n-r-1},
\]

where \( F_{\alpha;r+1, n-r-1} \) is the upper \( \alpha \) point of a \( F \)-variable with \( r+1 \) and \( n-r-1 \) degrees of freedom. This then gives

\[
P[|\frac{\sum \hat\beta_j u_j}{j=0^r} - \frac{\sum \hat\beta_j u_j}{j=0^r} u_0| \leq (r+1) F_{\alpha;r+1, n-r-1} \times ((u_0 u_1 \ldots u_r)(T'T)^{-1}(u_0 u_1 \ldots u_r)' \frac{1}{2} \gamma; (u_0 u_1 \ldots u_r) \in E_{r+1}) = 1-\alpha
\]
[6]. An infinite subset of the above intervals is then a confidence band of
confidence coefficient $\geq 1 - \alpha$ for the mean curve. For $r=1$ this gives a band
for the whole line with exact confidence coefficient $1-\alpha$. A little calculation
shows this to be the hyperbolic band referred to in Section 1.

3. STRAIGHT LINE REGRESSION

This section contains the analysis in detail of the straight line regression
case. For convenience the regression line is written in the form

$$m_t = \beta_0 + \beta_1 (t - \bar{t})$$

where $\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i$, $n > 2$. The $t_i$'s are observation points such that at least
two are distinct. The observation at $t_i$ is denoted by $y_i$. It is supposed that
observations may be made only in an interval $[A, B]$ and that a uniformly wide con-
fidence band is required for the interval $[a, b]$, i.e., $w(t) = 1$ for $t \in [a, b]$.

Proceeding as outlined in Section 2, form the function

$$\begin{align*}
\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} + \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} (t - \bar{t}),
\end{align*}$$

where

$$\begin{align*}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i, \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (t_i - \bar{t}) y_i}{\sum_{i=1}^{n} (t_i - \bar{t})^2}, \\
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 (t_i - \bar{t})]^2
\end{align*}$$
are stochastically independent. Determine for \( \delta > 0 \)

\[
V(-\delta, \delta) = \left\{ \left( \frac{\hat{\beta}_0 - \beta_0}{\delta}, \frac{\hat{\beta}_1 - \beta_1}{\delta} \right) : -\delta < \frac{\hat{\beta}_0 - \beta_0}{\delta} + \frac{\hat{\beta}_1 - \beta_1}{\delta} (t - \bar{t}) < \delta, t \in [a, b] \right\},
\]

or equivalently the image \( V'(\delta, \delta) \) of \( V(-\delta, \delta) \) in the plane of \( t \)-variables of \( n-2 \) degrees of freedom

\[
u = \sqrt{n} \frac{\hat{\beta}_0 - \beta_0}{\delta}, \quad \nu = \sqrt{ns} \frac{\hat{\beta}_1 - \beta_1}{\delta}
\]

where

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - \bar{t})^2.
\]

The resulting set is a parallelogram and is shown in Fig. 1. The density function of \( (u, \nu) \) is given

![Figure 1](image-url)
by

\[ g(u,v) = \frac{1}{2\pi} \left[ 1 + \frac{u^2 + v^2}{n-2} \right] - \frac{1}{2} n. \]

From symmetry of the density function we need to consider only the probability of the triangle \( T(-\delta, \delta) \) in the upper half plane.

An examination of Fig. 1 illustrates the fact that for a fixed \([a, b], [A, B], \delta, \text{ and } n\), different values of \( t \) and \( s \) result in different confidence coefficients.

The problem of maximizing the confidence coefficient is now investigated.

For a given \( t \) the claim is that the confidence coefficient is maximized when the variance of the observation points is maximized. For \( t \) such that the apex of the triangle is in \([-\sqrt{n}\delta, \sqrt{n}\delta]\), this is clear from the fact that if \( s_2^2 > s_1^2 \) are the variances of two configurations with corresponding triangles \( T_2(-\delta, \delta) \) and \( T_1(-\delta, \delta) \) respectively, then \( T_2(-\delta, \delta) \supset T_1(-\delta, \delta) \). If \( t \) is such that the apex of \( T(-\delta, \delta) \) lies in the complement of \([-\sqrt{n}\delta, \sqrt{n}\delta]\), it is not patently clear that the probability increases with \( s \). That this, however, is the case is shown as follows. Let \( h(\xi, \eta) = P[T(-\delta, \delta)] \), where \( \xi \) and \( \eta \) are the \( u \) and \( v \) coordinates of the apex. Then

\[
h(\xi, \eta) = \int_0^\eta \int_0^{\alpha_2(\xi, \eta)} \frac{1}{2\pi} \left[ 1 + \frac{u^2 + v^2}{n-2} \right] - \frac{1}{2} n, \]

where

\[
\alpha_1(\xi, \eta) = \frac{\xi + \sqrt{n}\delta}{\eta} v - \delta \sqrt{n},
\]

\[
\alpha_2(\xi, \eta) = \frac{\xi - \sqrt{n}\delta}{\eta} v + \delta \sqrt{n}.
\]
Hence

\[ 2\pi \frac{\partial}{\partial \eta} h(\xi, \eta) = \xi (I_1 - I_2) + 2 \sqrt{n \delta} (I_1 + I_2), \]

where

\[ I_1 = \int_0^\eta dv \left[ 1 + \frac{\alpha_1^2(\xi, \eta) + v^2}{n-2} \right] - \frac{n}{2} , \]

\[ I_2 = \int_0^\eta dv \left[ 1 + \frac{\alpha_2^2(\xi, \eta) + v^2}{n-2} \right] - \frac{n}{2} . \]

But \( I_1 \geq I_2 \) for \( \xi \geq 0 \) since

\[ \alpha_2^2(\xi, \eta) - \alpha_1^2(\xi, \eta) = \frac{4\xi \sqrt{n \delta} v}{\eta} \left[ 1 - \frac{v}{\eta} \right] \geq 0 , \quad 0 \leq v \leq \eta . \]

Thus for \( \xi \geq 0 \) (and by symmetry for \( \xi \leq 0 \))

\[ 2\pi \frac{\partial}{\partial \eta} h(\xi, \eta) \geq 0 . \]

This proves that for a given \( \bar{\tau} \), \( A \leq \bar{\tau} \leq B \), the variance of the \( n \) observation points must be maximized. Intuitively, this is what one would expect.

It can be shown (Appendix) that for any \( \bar{\tau} \), \( A \leq \bar{\tau} \leq B \), the corresponding maximum \( s^2 \) which may be attained by the observation points \( \{t_1, t_2, \ldots, t_n\} \) is

\[ (B-A)^2 \bar{\tau}^2 \left( \frac{k}{n} \right)^2 , \quad \frac{k}{n} \leq \tau \leq \frac{k+1}{n} , \quad k=0,1,\ldots,n-1 , \]

and
The configuration of observation points to obtain this maximum occurs with k \( t_i \)'s at B, l \( t_i \) at n(\( t_i - A \)) - k(B-A) + A, and n - (k+l) \( t_i \)'s at A. Thus for a fixed \( \bar{t} \), the maximum confidence coefficient for the band of width \( 2\sigma \) is achieved when the coordinates of the apex are

\[
\begin{align*}
  u &= 2 \sqrt{n} \ell (\bar{t} - e) , \\
  v &= \ell \sqrt{n} \ell f(\bar{t}) ,
\end{align*}
\]

where

\[
\ell = \frac{B-A}{b-a} ,
\]

and

\[
e = \frac{a+b}{2} - A .
\]

Plots of the loci of the apex are shown in Figures 2 and 3 for an even and an odd sample size respectively. Each section of the curve corresponds to the range

\[
\frac{k}{n} \leq \frac{\bar{t} - A}{B-A} \leq \frac{k+1}{n} , \quad k=0,1,...,n-1 .
\]
Due to the symmetry of the problem, it may always be assumed that \(-\infty < e \leq \frac{1}{2}\), i.e., the midpoint of the interval \([a, b]\) is always to the left of the midpoint of \([A, B]\). Whenever \(e = \frac{1}{2}\), i.e., \([A, B]\) and \([a, b]\) have the same midpoint. The
contours are symmetrical with respect to the \(v\)-axis. For \(e < \frac{1}{2}\), the contours are shifted to the right by the amount \(2\sqrt{n}\delta \left(\frac{1}{2} - e\right)\).

The problem of choosing the best \(t\) for a fixed \([A,B]\), \([a,b]\), \(n\), and \(\delta\) is now considered. The best \(t\) is defined as the one that yields the maximum confidence coefficient when its corresponding maximum variance configuration is used (or equivalently minimizes \(\delta\) for a given confidence coefficient \(1-\alpha\), \([A,B]\), \([a,b]\), and \(n\)). Intuitively one would expect the best \(t\) to be the one whose corresponding maximum variance configuration possesses the highest possible variance of the observation points. This has been proved for the following situations:

1. \(n\) even and \(\geq 6\), \(\frac{1}{2} \frac{n-2}{n-1} \leq e \leq \frac{1}{2}\). First it is shown that the maximum confidence coefficient must be attained for some point on the apex-contour-curve between the first peak to the left of the \(v\)-axis and the highest peak to the right of the \(v\)-axis. This follows from the fact that \(\frac{\partial}{\partial \eta} h(\xi, \eta) \geq 0\), and that

\[
2\pi \frac{\partial}{\partial \xi} h(\xi, \eta) = I_2 - I_1 \leq 0, \quad \xi \geq 0.
\]

This last equation merely states that the probability in the triangle decreases as its apex moves away from the \(v\)-axis along a horizontal line. Next observe that each section has an axis of symmetry for a distance (which may be 0, such as for the first and last sections) on either side of a vertical which passes through the point having a horizontal tangent (see the arc \(AB\), \(k=1\), in Fig. 2). Hence if the \(v\)-axis intersects any section to the right of the axis of symmetry, the maximum probability of a triangle whose apex lies anywhere on the section occurs when the apex is at or to the right of the \(v\)-axis. This is the situation
(see Fig. 2) when \( \frac{1}{2} \left( \frac{n-2}{n-1} \right) \leq e \leq \frac{1}{2} \), which means that the \( v \)-axis intersects the first section somewhere on the arc CE.

Now as the apex moves from the \( v \)-axis toward the peak, the probability in the triangle, which is one half the confidence coefficient, increases. This follows by writing the confidence coefficient as a function of \( \tau = \frac{\ell-A}{B-A} \) in the iterated integral

\[
P[V'(-\delta, \delta)] = 2 \int_0^\infty dv \int_{\nu_1(\tau)}^{\nu_2(\tau)} du \frac{1}{2\pi} \left[ 1 + \frac{u^2 + v^2}{n-2} \right] - \frac{n}{2},
\]

where

\[
\nu_1(\tau) = \frac{2\ell(\tau-e) + 1}{2\ell(\tau)} v - \sqrt{n}\delta,
\]

\[
\nu_2(\tau) = \frac{2\ell(\tau-e) - 1}{2\ell(\tau)} v + \sqrt{n}\delta,
\]

and

\[
\phi(\tau) = 2 \sqrt{n}\delta \ell(\tau).
\]

Differentiating with respect to \( \tau \) gives

\[
2\pi\ell^2(\tau) \frac{\partial}{\partial \tau} P[V'(-\delta, \delta)] = \frac{2\ell}{f(\tau)} [\tau ((n-1)e-k) + \frac{k}{n} (1 + \frac{k}{n}) - ke] (J_2 - J_1) + f'(\tau)(J_1 + J_2)
\]

where

\[
J_1 = \int_0^\infty dv v \left[ 1 + \frac{\nu_1^2(\tau) + v^2}{n-2} \right] - \frac{n}{2},
\]
As the apex moves from the v-axis toward the peak (along arc DE on Fig. 2), \( \tau \) varies from \( e \) to \( \frac{1}{2} \), \( k = \frac{n}{2} - 1 \), \( f'(\tau) > 0 \), and \( J_2 - J_1 < 0 \). But for \( n=6,8,10,... \) the coefficient of \( J_2 - J_1 \) is < 0 along the arc DE and hence \( \frac{\partial}{\partial \tau} P[V'(-5,8)] > 0 \). This means that the maximum confidence coefficient is attained when the apex of the triangle is at the point E.

2. \( n \) odd and \( \geq 3 \), \( \frac{n-1}{2n} \leq e \leq \frac{1}{2} \). In this case the v-axis lies somewhere on arc EG, say F, Fig. 3. The maximum probability is then on arc EF. As the apex moves from E to F, \( \tau \) varies from \( \frac{n-1}{2n} \) to \( e \), \( k = \frac{n-1}{2} \), \( f'(\tau) < 0 \), and \( J_2 - J_1 > 0 \). But the coefficient of \( J_2 - J_1 \) is < 0 along EG and hence \( \frac{\partial}{\partial \tau} P[V'(-5,8)] < 0 \). Thus the maximum confidence coefficient occurs when the apex of the triangle is at E.

3. \( n \) odd and \( \geq 7 \), \( \frac{n-3}{2(n-1)} \leq e \leq \frac{n-1}{2n} \). Now the v-axis would lie on CE, say D, in Fig. 3, and the maximum probability would lie somewhere on arc DE. As the apex moves from D to E, \( \tau \) varies from e to \( \frac{n-1}{2n} \), \( k = \frac{n-3}{2} \), \( f'(\tau) > 0 \), and \( J_2 - J_1 \leq 0 \). But the coefficient of \( J_2 - J_1 \) is < 0 along DE and \( \frac{\partial}{\partial \tau} P[V'(-5,8)] > 0 \), i.e., the maximum confidence coefficient occurs at E.

4. **TABLE**

From the above it is seen that in general, the maximum confidence coefficient for a band of width \( 2\delta \sigma \) depends on the parameters \( c = \frac{B-A}{b-a} \), \( e = \frac{a+b}{B-A} \), and \( n \). Hence, a table which could handle all possible experimental situations would
have to contain the value of the confidence coefficient for a range of values of the parameters $\delta$, $l$, $e$, and $n$. This seemed too extensive an undertaking at this time.

The table presented in this paper is constructed for use in the following situation:

(1) $n$ even, specifically, $n = 4(2)20(10)30(20)50, \infty$.

(2) $\tau = \frac{a+b}{2}$, so that an optimum solution is possible only if $\frac{a+b}{2} = \frac{A+B}{2}$.

Hence the problem is essentially to compute the integral of the function

$$g(u,v) = \frac{1}{2\pi} \left[ 1 + \frac{u^2 + v^2}{n-2} \right] - \frac{1}{2} n$$

over the triangle shown in Fig. 4.
The table consists of 13 pages. At the top of each page are listed two values of a number \( c = 1.1.2.2.3.4.5.6.10.10.20, \infty \). When the maximum variance configuration is used, i.e., \( \frac{n}{2} \) observations at A and \( \frac{n}{2} \) observations at B, \( c = \frac{b-A}{b-a} = \ell \). If any other configuration of observation points is used, still maintaining \( \ell = \frac{a+b}{2} \), then \( c = \frac{2s}{b-a} \) where \( s \) is the variance of the observation points. For each value of \( c \), the confidence coefficient is computed for all combinations of \( n = 4(2)20(10)30(20)50, \infty \) and \( d = \sqrt{n}\delta = 1(.05)2.5(.1)4(.2)5(.5)7(1)10(5)20(10)50 \). The confidence coefficient is entered into the body of the table without a decimal point. Each entry is correct to 3 significant figures and a blank space corresponds to a rounding off to 1.

It should be noted that the table is not restricted to those values of \( c \geq 1 \). Because of the symmetry of the density function \( g \), it follows that for any \( c < 1 \), the table with heading \( 1/c \) may be used. In this case the values in the column \( d = \sqrt{n}\delta \) must be multiplied by \( 1/c \).

This table was computed using an expression derived by a technique similar to that of Dunnett and Sobel [5]. The confidence coefficient \( 1-\alpha \) may be written as

\[
\frac{1}{4}(1-\alpha) = \frac{n-2}{2\pi} \int_{0}^{\pi/2} d\theta \int_{0}^{\infty} r(\theta) \int_{0}^{(1+\theta^{2})^{-1/2}} \frac{n}{2} - 1 \n\]

\[
= \frac{1}{4} - \frac{1}{2\pi} \int_{\tan^{-1}c}^{\pi/2} d\phi [1 + k^{2} \csc^{2}\phi]^{1/2} \frac{n}{2} + 1
\]

*The table is composed of computer print-out and thus the letters \( c, n, \) and \( d \) appear as capitals.
where

\[ \phi^2 = \frac{u^2}{n-2} + \frac{v^2}{n-2} = \bar{\rho} \sqrt{\frac{n}{n-2}} \frac{\sin \psi}{\sin(\theta + \psi)}, \]

\[ \theta = \tan^{-1} \frac{v}{u}, \]

\[ \psi = \tan^{-1} c, \]

\[ k^2 = \frac{\delta^2 c^2 n}{(n-2)(1+c^2)}, \]

\[ c = \frac{2s}{b-a}, \]

\[ \varphi = \theta + \psi. \]

Define

\[ Q_n = \frac{1}{2\pi} \int_{\tan^{-1} c}^{\frac{\pi}{2} + \tan^{-1} c} d\phi [1 + k^2 \csc^2 \varphi] - \frac{n}{2} + 1 \]

and consider

\[ Q_n - Q_n = -\frac{1}{2\pi} \int_{\tan^{-1} c}^{\frac{\pi}{2} + \tan^{-1} c} d\phi [1 + k^2 \csc^2 \varphi] - \frac{n}{2} + 1 + k^2 \csc^2 \varphi. \]

Making use of the change of variable

\[ y = \frac{1}{1 + \left(1 + \frac{1}{k^2}\right) \tan^2 \varphi} \]

it is seen after some calculation that
\[
\frac{Q_n}{2} - \frac{Q_n}{2} - 1 = \frac{k(l+k)^2 - \frac{1}{2}(n-3)}{4\pi} \left\{ B_{\frac{1}{2}}(c,k) \left[ \frac{1}{2}, \frac{1}{2}(n-3) \right] + B_{\frac{1}{2}}(c,k) \left[ \frac{1}{2}, \frac{1}{2}(n-3) \right] \right\}
\]

where

\[
f_1(c,k) = \frac{1}{1 + \left(1 + \frac{1}{k^2}\right)c^2},
\]

\[
f_2(c,k) = \frac{1}{1 + \left(1 + \frac{1}{k^2}\right)\frac{1}{c^2}},
\]

and \( B_z[p,q] = \int_0^1 t^{p-1}(1-t)^{q-1} dt \) is the incomplete beta function.

Now for the case that \( n \) is odd and \( \geq 3 \)

\[
1 - \alpha = 1 - 4Q_n/2
\]

\[
= 1 - 4\left[ (Q_{\frac{n}{2}} - Q_{\frac{n}{2}} - 1) + (Q_{\frac{n}{2}} - Q_{\frac{n}{2}} - 2) + \ldots + (Q_{\frac{n}{2}} - Q_{\frac{n}{2}} + Q_{\frac{n}{2}}) \right].
\]

But

\[
Q_{\frac{3}{2}} = \frac{1}{2\pi} \left[ \sin^{-1} \frac{1}{\sqrt{1+(1+n8^2)c^2}} + \sin^{-1} \frac{c}{\sqrt{1+(1+n8^2)c^2}} \right]
\]

so that finally, in terms of the incomplete beta function ratio \( I_z[p,q] = B_z[p,q]/B_1[p,q] \).
17 April 1963

\[ 1 - \alpha = 1 - \frac{2}{\pi} \left[ \sin^{-1} \frac{1}{\sqrt{1+(1+n^2)\beta^2}} + \sin^{-1} \frac{c}{\sqrt{1+(1+n^2)\beta^2}} \right] \]

\[ + \frac{2k}{\pi} \left[ \frac{1}{2} \sum_{j=1}^{n-3} \frac{4^{j-1}[(j-1)!]^2}{(1+k^2)^j(2j-1)!} \left( I_{f_1}(c,k) \left[ \frac{1}{2}, j \right] + I_{f_2}(c,k) \left[ \frac{1}{2}, j \right] \right) \right], \quad n=5,7,9,... \]

\[ = 1 - \frac{2}{\pi} \left[ \sin^{-1} \frac{1}{\sqrt{1+(1+n^2)\beta^2}} + \sin^{-1} \frac{c}{\sqrt{1+(1+n^2)\beta^2}} \right], \quad n=3. \]

The formula

\[ I\left( \frac{1}{2}, j \right) = \sqrt{z} \sum_{i=0}^{n-1} \frac{(2i)!}{4^i(i!)^2} (1-z^i) \]

is used for calculating the incomplete beta function ratios in (1).

For \( n \) even and \( \geq 4 \)

\[ 1 - \alpha = 1 - 4 \frac{Q_n}{2} \]

\[ = 1 - 4 \left[ \left( \frac{Q_n}{2} - \frac{Q_{n-2}}{2} - 1 \right) + \left( \frac{Q_n}{2} - \frac{Q_{n-2}}{2} - 1 \right) + \ldots + \left( Q_2 - Q_2 + Q_2 \right) \right]. \]

But \( Q_1 = \frac{1}{4} \). Hence after some calculation

\[ 1 - \alpha = k \sum_{j=1}^{n/2-1} \frac{(2j-1)!}{(1+k^2)^j \frac{1}{2} [(j-1)!]^2 4^{j-1}} \left( I_{f_1}(c,k) \left[ \frac{1}{2}, j \right] + I_{f_2}(c,k) \left[ \frac{1}{2}, j \right] \right) \]

\[ n=4,6,8,... . \]
The formula

$$I_z\left(\frac{1}{2}, j \cdot \frac{1}{2}\right) = \frac{2}{\pi} \tan^{-1} \sqrt{\frac{z}{1-z}} + \frac{2}{\pi} \sqrt{z} (1-z) \sum_{i=0}^{\frac{1}{2}} \frac{4^i (i+1)^2}{(2i+1) (1-z)^{i+1}}$$

is used for evaluating the incomplete beta function ratios appearing in (2).

The actual computations were performed on a Philco 2000 digital computer using equations (1) and (2). For $n \leq 50$, which is the range of finite $n$ in the table, an error analysis showed that the resulting probabilities could be off at most by seven digits in the 7th place. To reduce the size of the table, however, these were rounded off to three figures. This should be sufficient for most applications.

Now

$$\lim_{n \to \infty} g(u, v) = \frac{1}{2\pi} e^{-\frac{1}{2}(u^2 + v^2)}$$

which is the uncorrelated bivariate normal distribution with zero means and unit variances. To make the calculation for $n=\infty$, which amounts to the integral of (3) over the triangle of Fig. 4, a method outlined by Owen [7] was used. For $1 \leq c < \infty$ this gives

$$1 - \alpha = 1 - 4(E + F)$$

where

$$E = T\left(x, \frac{1}{c}\right),$$
\[ F = \frac{1}{2} \left[ G(x) + G(y) \right] - \left[ G(x)G(y) + \Pi \left( y, \frac{1}{c} \right) \right], \]

\[ x = \frac{c(5\sqrt{n})}{\sqrt{1+c^2}}, \]

\[ y = \frac{c^2(5\sqrt{n})}{\sqrt{1+c^2}}, \]

\[ T(h, z) = \frac{1}{2\pi} \left( \tan^{-1} z - \sum_{j=0}^{\infty} c_j z^{2j+1} \right), \]

\[ c_j = (-1)^j \frac{1}{2j+1} \left[ 1 - e^{-\frac{h^2}{2}} \sum_{i=0}^{j} \frac{(\frac{h^2}{2})^i}{i!} \right], \]

\[ G(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{1}{2} \xi^2} d\xi. \]

For \( c=\infty, E=0 \) and

\[ F = \frac{1}{2} [1 - G(x)] \]

where

\[ x = (5\sqrt{n}). \]

These again were performed on the Philco 2000 and the computations were such that the resulting confidence coefficients are correct to three significant figures.

One additional observation is that, as \( c=\infty \) for a fixed \( 5\sqrt{n} \), the confidence
coefficient is the area of the function \( g \) over an infinite strip parallel to the \( v \)-axis. Hence, the values in the table with \( c^{\infty} \) could have been obtained from a \( t \)-table. Each column corresponds to a \( t \)-variable whose degrees of freedom is two less than the sample size heading.

5. **EFFICIENCY**

In Scheffe [3, pp. 52, 53], it is seen that a \( 1-\alpha \) confidence band for the true line consists of all points \((t, y)\) satisfying

\[
[y - \hat{\alpha} - \hat{\beta}(t - \bar{t})]^2 \leq F_{\alpha; 2, n-2} \hat{\sigma} \left[ \frac{1}{n} + \frac{(t - \bar{t})^2}{ns^2} \right].
\]

This gives a band about the fitted line, bounded by the two branches of a hyperbola. In order to use this for comparison purposes with the method of this paper, it is restricted to just the interval \([a, b]\). The confidence coefficient of this band is, of course, no longer \( 1-\alpha \) but \( \geq 1-\alpha \).

The area \( A_1 \) of the hyperbolic band over the interval \([a, b]\) is given by

\[
A_1 = 2\hat{\sigma} \sqrt{2F_{\alpha; 2, n-2}} \int_a^b \frac{1}{n} + \frac{(t - \bar{t})^2}{ns^2} \frac{1}{2}.
\]

It is clear that this area is minimized when \( \bar{t} = \frac{a+b}{2} \) and \( s^2 \) is maximized. Thus if \( \frac{a+b}{2} = \frac{A+B}{2} \), \( s^2 \) is maximized for \( n \) even when \( \frac{n}{2} \) observations are at \( A \) and \( \frac{n}{2} \) observations are at \( B \). In this case \( s^2 = \frac{1}{4}(B-A)^2 \). Thus
\[ A_1 = \hat{\sigma}(b-a) c \sqrt{\frac{2F_{\alpha;2,n-2}}{n}} \int_0^c \frac{1}{(1+\xi^2)^{1/2}} d\xi \]

\[ = \hat{\sigma}(b-a) \sqrt{\frac{2F_{\alpha;2,n-2}}{n}} \left[ \left( 1 + \frac{1}{c^2} \right)^{1/2} + c \log \frac{\sqrt{1+c^2+1}}{c} \right], \]

where \( c = \frac{b-A}{b-a} \). The area \( A_2 \) of our band is \( 25\hat{\sigma}(b-a) \). Hence, the ratio \( A_1/A_2 \), which will be referred to as the efficiency of our method, is given by

\[ \frac{A_1}{A_2} = \frac{\sqrt{2F_{\alpha;2,n-2}}}{(5\sqrt{n})} \left[ \left( 1 + \frac{1}{c^2} \right)^{1/2} + c \log \frac{\sqrt{1+c^2+1}}{c} \right]. \] (4)

Eq. (4) is valid for any \( 0 < c < \infty \). It was noted in Section 4 that \( \lim_{c \to \infty} (5\sqrt{n}) = t_{\alpha/2,n-2} \).

But

\[ \lim_{c \to \infty} c \log \frac{\sqrt{1+c^2+1}}{c} = 1. \]

Hence

\[ \lim_{c \to \infty} \frac{A_1}{A_2} = \frac{\sqrt{2F_{\alpha;2,n-2}}}{t_{\alpha/2,n-2}}. \] (5)

The symmetry of the function \( g \) means that \( \lim_{c \to \infty} c\sqrt{n} = t_{\alpha/2,n-2} \). Also

\[ \lim_{c \to \infty} c^2 \log \frac{\sqrt{c^2+1}+1}{c} = 0 \]

Hence
Eqs. (4), (5), and (6) summarize the results of this section. Fig. 5 is a graph of the efficiency, for each of three values of n, as a function of c. The confidence coefficient selected is .95.
APPENDIX

Let \( \{z_1, z_2, \ldots, z_n\} \) be \( n \) points in the unit interval \([0,1]\). For a fixed \( z = \frac{1}{n} \sum_{i=1}^{n} z_i \) in \([0,1]\) the problem is to maximize

\[
s^2 = \sum_{i=1}^{n} (z_i - z)^2 = \sum_{i=1}^{n} z_i^2 - nz^2.
\]

The claim is that, to maximize set each \( z_i \) equal to 0 or to 1, except for one, keeping \( \sum_{i=1}^{n} z_i = nz \). For suppose, without loss of generality, \( 0 < z_1 \leq z_2 < 1 \). Then there exists \( \delta > 0 \) such that

\[
0 \leq z_1 - \delta \leq 1,
\]
\[
0 \leq z_2 + \delta \leq 1,
\]

and

\[
(z_1 - \delta)^2 + (z_2 + \delta)^2 = z_1^2 + z_2^2 + 2\delta^2 + 2\delta(z_2 - z_1) > z_1^2 + z_2^2,
\]

i.e., \( ns^2 \) may be increased. The actual configuration for any \( z \) such that \( \frac{k}{n} \leq z \leq \frac{k+1}{n} \), \( k = 0, 1, \ldots, n-1 \) is \( k \) \( z_i \)'s at 1, 1 \( z_i \) at \( nz-k \), and \( n-(k+1) \) \( z_i \)'s at 0.

The resulting maximum variance is

\[
\frac{k+(nz-k)}{n}^2 - \frac{z^2}{2}.
\]

Thus for \( \{t_1, t_2, \ldots, t_n\} \subseteq [A,B] \), the maximum variance configuration for a fixed \( t_i \), such that
\[
\frac{k}{n} \leq \tau = \frac{t-A}{B-A} \leq \frac{k+1}{n}, \quad k=0,1,\ldots,n-1,
\]

is given by \( k \ t_i \)'s at \( B \), \( 1 \ t_i \) at \( n(t-A) - k(B-A) + A \), and \( n-(k+1) \ t_i \)'s at \( A \).

The maximum variance is

\[
(B-A)^2 \left[ \frac{k+(n\tau-k)^2}{n} - \tau^2 \right].
\]
REFERENCES


<table>
<thead>
<tr>
<th>C</th>
<th>N</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
<th>22</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
<th>32</th>
<th>34</th>
<th>36</th>
<th>38</th>
<th>40</th>
<th>42</th>
<th>44</th>
<th>46</th>
<th>48</th>
<th>50</th>
<th>51N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>2.00</td>
<td>3.00</td>
<td>4.00</td>
<td>5.00</td>
<td>6.00</td>
<td>7.00</td>
<td>8.00</td>
<td>9.00</td>
<td>10.00</td>
<td>11.00</td>
<td>12.00</td>
<td>13.00</td>
<td>14.00</td>
<td>15.00</td>
<td>16.00</td>
<td>17.00</td>
<td>18.00</td>
<td>19.00</td>
<td>20.00</td>
<td>21.00</td>
<td>22.00</td>
<td>23.00</td>
<td>24.00</td>
<td>25.00</td>
<td>26.00</td>
<td>27.00</td>
</tr>
</tbody>
</table>

**C = 1.00**

<table>
<thead>
<tr>
<th>C</th>
<th>N</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
<th>22</th>
<th>24</th>
<th>26</th>
<th>28</th>
<th>30</th>
<th>32</th>
<th>34</th>
<th>36</th>
<th>38</th>
<th>40</th>
<th>42</th>
<th>44</th>
<th>46</th>
<th>48</th>
<th>50</th>
<th>51N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>2.00</td>
<td>3.00</td>
<td>4.00</td>
<td>5.00</td>
<td>6.00</td>
<td>7.00</td>
<td>8.00</td>
<td>9.00</td>
<td>10.00</td>
<td>11.00</td>
<td>12.00</td>
<td>13.00</td>
<td>14.00</td>
<td>15.00</td>
<td>16.00</td>
<td>17.00</td>
<td>18.00</td>
<td>19.00</td>
<td>20.00</td>
<td>21.00</td>
<td>22.00</td>
<td>23.00</td>
<td>24.00</td>
<td>25.00</td>
<td>26.00</td>
<td>27.00</td>
</tr>
</tbody>
</table>

**C = 1.1**

**C = 1.11**

17 April 1963

SP-1181/000/00
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
<tr>
<td>1.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
<tr>
<td>1.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
</tbody>
</table>

**Table:**

- **Columns:** 
  - Date
  - 1963

**Notes:**

- The table appears to be a calendar or schedule for the year 1963, with dates and corresponding numbers, possibly indicating events or data entries for each day.
System Development Corporation, Santa Monica, California

CONFIDENCE BANDS IN STRAIGHT LINE REGRESSION.

Scientific rept., SP-1181/000/00, by A. V. Gafarian. 17 April 1963, 45p., 7 refs., 5 figs.

Unclassified report

DESCRIPTORS: Statistical Distribution.
Statistical Functions.

Develops a method for obtaining confidence bands in polynomial regression when the observations are independently distributed with constant but unknown variance. Reports that the bands may be obtained over arbitrary sets of the independent variable with exact preassigned confident coefficients. Also reports that difficult distribution problems result when specific applications are attempted. Discusses first degree polynomials since some progress has been made here. Provides a table that obtains a constant width confidence band which contains the true but unknown straight regression line for values of the independent variable in some arbitrarily selected interval with an exact preassigned confident coefficient. Compares the present method with the classical hyperbolic band for the whole regression line.