CONFIDENCE INTERVALS FOR HERITABILITY IN A MIXED LINEAR MODEL

Brent D. Burch

Hari K. Iyer

CONFIDENCE INTERVALS FOR HERITABILITY IN A MIXED LINEAR MODEL

Brent D. Burch[1] and Hari K. Iyer[2]

[1]Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, Arizona 86011, U.S.A.

[2]Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.

ABSTRACT

A family of procedures is given to construct confidence intervals for the heritability of a trait in a mixed linear model. The procedures are applicable for constructing confidence intervals for a ratio of variance components in a mixed linear model having two sources of variation. The resulting intervals are evaluated in terms of expected length. The investigator may select the best confidence interval procedure from the family of procedures based on the interval(s) having short expected length. Confidence intervals for loineye data using bulls from a Red Angus seed stock herd will be presented.

1 Introduction

Linear models are frequently used in applications such as plant and animal breeding. In these fields of study the components of variation are often perceived as having either a genetic or environmental (non-genetic) origin. Mixed linear models (models which take into account both fixed and random effects) having two variance components are often used with the random effect corresponding to the genetic source and the error corresponding to the environmental source. In many cases it is assumed that the genetic and environmental effects are independent. However, the observational units are not independent of one another if they possess common genetic material. In other words, the covariance between two observations will usually involve the genetic component of the overall variance. Additionally, the random effects themselves may be correlated with one another depending on the genetic relationship between the corresponding observational units.

In plant and animal breeding, inferences concerning functions of variance components are often of primary importance. We define the ratio of variance components, denoted by γ, as the genetic component of variance divided by the environmental component of variance. It follows that $\rho = \gamma/(1+\gamma)$ is the proportion of total variation due to genetic effects. ρ is sometimes referred to as the heritability of the trait under study.

One of the complexities in applying mixed linear models to plant and animal breeding problems is accurately modeling the covariances among observations. The approach used to accomplish this task is to obtain a matrix that describes the (additive) genetic relationship

between the observations. This matrix, denoted by $\boldsymbol{A}$, is called the relationship matrix. Henderson (1976) devised a recursive method for computing $\boldsymbol{A}$ as well as a rapid method for computing the inverse of $\boldsymbol{A}$.

The most common method of developing confidence intervals is the pivotal quantity technique. This approach centers on finding a function of the data and the parameter of interest, say, ρ, whose distribution is free of all model parameters. If this quantity is a monotone function of ρ, a confidence interval for ρ can be obtained. There are cases in which numerous pivotal quantities are available that result in exact confidence intervals for ρ. LaMotte et al. (1988) discussed the use of these quantities in the context of hypothesis tests.

The paper is organized as follows. In Section 2 we provide an overview of a general mixed linear model having two variance components. The variance components are denoted by σ_a^2 and σ_e^2 where σ_a^2 is the "additive" component of genetic variance and σ_e^2 is the "environmental" variance component. The structure of the model, observable random variables, unobservable random variables, unknown parameters, and distributional assumptions are discussed. In our case, the parameters under study are the variance components. With this in mind, minimal sufficient statistics are derived for the model void of the location parameters.

The particular function of variance components we are interested in is $\rho = \sigma_a^2/(\sigma_a^2+\sigma_e^2)$. Confidence intervals for ρ are discussed and we employ the pivotal quantity method to obtain these confidence intervals. Due to the multitude of possible pivotal quantities, criteria of good confidence intervals are entertained. Expected length is a traditional measure of goodness and is the one used in this paper. In fact, we investigate the expected length properties of confidence intervals for ρ as developed from the pivotal quantities considered by LaMotte and McWhorter (1978) for mixed linear models having two variance components. Expected length computations are made possible by a result given by Pratt (1961).

In Section 3 we present data that consists of measurements on one hundred and seventy one yearling bulls from a Red Angus seed stock herd in Montana (Evans et al. (1995)). We demonstrate that it is possible to numerically compute expected lengths for confidence intervals for moderately sized data sets. For this data set, there are one hundred and sixty four pivotal quantities that may be inverted to obtain confidence intervals for ρ. The expected length of 90% confidence intervals for a few of these pivotal quantities are given as a function of the parameter. In Section 4 we provide a summary of the paper and discuss how one may use the expected length results to make decisions related to confidence interval construction.

2 Mixed Linear Model

The mixed linear model under consideration is

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of observable random variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{u}$ and $\boldsymbol{e}$ are vectors of unobservable random variables of size $m \times 1$ and $n \times 1$, respectively. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known and without loss of generality, rank($\boldsymbol{X}$) $= p$. It is assumed that $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent where $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{A})$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_n)$. In animal breeding contexts, the known matrix $\boldsymbol{A}$ is referred to as the relationship matrix since it describes the degree to which the $\boldsymbol{u}$'s are related. In that scenario, if the elements u_1 and u_2 of $\boldsymbol{u}$ are the (additive) genetic effects corresponding to a parent and offspring, respectively, then Cov(u_1, u_2) $= \sigma_a^2/2$ (see Falconer, 1989, p.150). It follows that $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_e^2 \boldsymbol{I}_n + \sigma_a^2 \boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}')$.

We now illustrate how the mixed linear model notation is used in a simple example. Consider a situation in which there are five animals under study. These animals reside at two different ranches and the relationships among the animals are depicted in Figure 1. Animals 1,3, and 4 belong to ranch 1 and animals 2 and 5 belong to ranch 2.

Let Y_i be the response of the i^{th} animal, $i = 1, ..., 5$. Then $\boldsymbol{Y}' = [Y_1, Y_2, Y_3, Y_4, Y_5]$ is the vector of responses for the animals. Taking into account the location (i.e., ranch) effect and (additive) genetic effect on the observations, the mixed linear model may be expressed as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}. \tag{2}$$

From Figure 1, animals 1, 3, and 4 are related and animals 2 and 5 are related. The relationship matrix $\boldsymbol{A}$ for the example is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & .5 & .5 & 0 \\ & 1 & 0 & 0 & .5 \\ & & 1 & .25 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix}. \tag{3}$$

Since animal 3 and animal 4 are offspring of animal 1, and animal 5 is an offspring of animal 2, Cov(u_1, u_3) $= \sigma_a^2/2$, Cov(u_1, u_4) $= \sigma_a^2/2$, and Cov(u_2, u_5) $= \sigma_a^2/2$. Also note that animals 3 and 4 are half-sibs. That is, they have one parent in common and the other parent is different. It follows that Cov(u_3, u_4) $= \sigma_a^2/4$.

In order to construct confidence intervals for the heritability of a trait, it is advantageous to find the statistics that are useful in estimating the variance components. Using the notation given in Section 1, $\gamma = \sigma_a^2/\sigma_e^2$. In the usual manner, we take $\sigma_a^2 \geq 0, \sigma_e^2 > 0$ so that $0 \leq \gamma < \infty$ and $0 \leq \rho < 1$. Let $\boldsymbol{H}$ be a $n \times (n - p)$ matrix whose columns span the space orthogonal to the space spanned by the columns of $\boldsymbol{X}$ and satisfy $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_{n-p}$. Then

$\boldsymbol{H'Y} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_{n-p} + \sigma_a^2 \boldsymbol{H'ZAZ'H})$. Let $0 \leq \Delta_1 < ... < \Delta_d$ be the distinct eigenvalues of $\boldsymbol{H'ZAZ'H}$ having multiplicities $r_1, ..., r_d$, respectively. There exists an $(n-p) \times (n-p)$ orthogonal matrix $\boldsymbol{P}$ such that $\boldsymbol{P'(H'ZAZ'H)P} = Diag(\Delta_1, ..., \Delta_1, ..., \Delta_d, ..., \Delta_d)$ where each Δ_i is repeated r_i times, $i = 1, ..., d$. It follows that $\boldsymbol{H'ZAZ'H} = \sum_{i=1}^{d} \Delta_i \boldsymbol{P}_i \boldsymbol{P}_i'$ where $\boldsymbol{P} = [\boldsymbol{P}_1, ..., \boldsymbol{P}_d]$ and each matrix $\boldsymbol{P}_i$ corresponding to Δ_i is of size $(n-p) \times r_i$. For $i = 1, ..., d$, $\boldsymbol{P}_i'\boldsymbol{H'Y} \sim N(\boldsymbol{0}, (\sigma_e^2 + \sigma_a^2 \Delta_i)\boldsymbol{I}_{r_i})$. So $\boldsymbol{Y}'(H\boldsymbol{P}_i\boldsymbol{P}_i'H')Y = Q_i \sim (\sigma_e^2 + \sigma_a^2\Delta_i)\chi^2(r_i)$, $i = 1, ..., d$. By construction, the quadratic forms $Q_1, ..., Q_d$ are independent. In addition, they are a set of minimal sufficient statistics associated with the reduced linear model void of the fixed effect. Rewriting the distribution of Q_i in terms of ρ and the nuisance parameter σ_e^2, we have that $Q_i \sim \sigma_e^2(1 + \Delta_i\rho/(1-\rho))\chi^2(r_i)$, $i = 1, ..., d$. These quadratic forms play a central role in the construction of the confidence intervals.

Using the results of LaMotte and McWhorter (1978), a pivotal quantity that can be inverted to obtain confidence intervals for ρ and its associated distribution is

$$G_k(\rho) = \frac{\sum_{i=k+1}^{d} \frac{Q_i}{1+\rho(\Delta_i - 1)} \Big/ \sum_{i=k+1}^{d} r_i}{\sum_{i=1}^{k} \frac{Q_i}{1+\rho(\Delta_i-1)} \Big/ \sum_{i=1}^{k} r_i} \sim F\left(\sum_{i=k+1}^{d} r_i, \sum_{i=1}^{k} r_i\right) \tag{4}$$

where k ranges from 1 to $d-1$. Since there are d distinct eigenvalues, there are $d-1$ possible pivotal quantities that are monotone decreasing functions in ρ. These quantities can be inverted numerically to obtain confidence intervals for ρ. Notationally, a 100(1-α)% confidence interval for ρ is given by the set

$$\left\{ \rho \in [0,1) : F_{\alpha_1} \leq \frac{\sum_{i=k+1}^{d} \frac{Q_i}{1+\rho(\Delta_i - 1)} \Big/ \sum_{i=k+1}^{d} r_i}{\sum_{i=1}^{k} \frac{Q_i}{1+\rho(\Delta_i-1)} \Big/ \sum_{i=1}^{k} r_i} \leq F_{1-\alpha_2} \right\} \tag{5}$$

where $\alpha_1 + \alpha_2 = \alpha$ and $F_{\alpha_1}, F_{1-\alpha_2}$ are the $\alpha_1, 1-\alpha_2$ percentiles of the F distribution having numerator and denominator degrees of freedom equal to $\sum_{i=k+1}^{d} r_i$ and $\sum_{i=1}^{k} r_i$, respectively. Let L denote the infimum of this set and U the supremum. Then $P[L \leq \rho \leq U] = 1 - \alpha$. Figure 2 depicts the process of inverting $G_k(\rho)$ to obtain a confidence interval for ρ. The probability that the pivotal quantity $G_k(\rho)$ is between F_{α_1} and $F_{1-\alpha_2}$ is equal to the probability that ρ is between L and U.

The confidence intervals obtained by inverting the pivotal quantity in (4) depend on the quadratic forms $Q_1, ..., Q_d$. For $\boldsymbol{Q} = (Q_1, ..., Q_d)$, the expected length of a confidence interval $[L, U] = [L(\boldsymbol{Q}), U(\boldsymbol{Q})]$ is

$$E[U(\boldsymbol{Q}) - L(\boldsymbol{Q})] = \int \cdots \int [U(\boldsymbol{Q}) - L(\boldsymbol{Q})] f_{\boldsymbol{Q}}(\boldsymbol{q}) d\boldsymbol{q} \tag{6}$$

where $f_{\boldsymbol{Q}}$ is the d-dimensional probability density function of $\boldsymbol{Q}$. For large values of d, evaluating the expectation in (6) by direct numerical integration techniques may be

computationally infeasible. An alternative expression for expected length is

$$E\left[U(\boldsymbol{Q})\right] - E\left[L(\boldsymbol{Q})\right] \;=\; \int v f_V(v)dv - \int w f_W(w)dw \tag{7}$$

where $V = U(\boldsymbol{Q})$, $W = L(\boldsymbol{Q})$, and f_V, f_W are the one dimensional probability density functions of V, W, respectively. Using (7) appears intractable since the distributions of the endpoints of the confidence interval are not known.

Pratt (1961) derived a relationship between the expected length of a confidence interval and the probability of false coverage. That is,

$$E_{\rho_T}\left[U(\boldsymbol{Q}) - L(\boldsymbol{Q})\right] \;=\; \int\limits_{\rho \neq \rho_T} P_{\rho_T}\left[L \leq \rho \leq U\right] d\rho \tag{8}$$

where ρ_T is the true value of the parameter and $P_{\rho_T}\left[L \leq \rho \leq U\right]$ is the probability that the confidence interval covers an arbitrary value of ρ. Note that $P_{\rho_T}\left[L \leq \rho_T \leq U\right] = 1 - \alpha$. The subscripts in (8) accentuate the fact that the probability and expectation are functions of ρ_T. Using notation involving the quadratic forms $Q_1, ..., Q_d$, the expression for expected length becomes

$$E_{\rho_T}\left[U(\boldsymbol{Q}) - L(\boldsymbol{Q})\right] \;=\; \int\limits_0^1 P_{\rho_T}\left[F_{\alpha_1} \leq \frac{\sum\limits_{i=k+1}^{d} \frac{Q_i}{1+\rho(\Delta_i - 1)} \Big/ \sum\limits_{i=k+1}^{d} r_i}{\sum\limits_{i=1}^{k} \frac{Q_i}{1+\rho(\Delta_i - 1)} \Big/ \sum\limits_{i=1}^{k} r_i} \leq F_{1-\alpha_2}\right] d\rho \tag{9}$$

which can be computed numerically since the integrand in (9) can be written as the difference in probabilities of linear combinations of independent chi-square random variables. Based on the work of Imhoff (1961), Davies (1980) developed an algorithm that computes the cumulative distribution function of a linear combination of independent chi-square random variables. Our program to compute expected lengths of confidence intervals uses Davies' FORTRAN routine which is also known as algorithm AS 155 from *Applied Statistics*.

Inverting the pivotal quantity in (4) may result in values of ρ that are outside the parameter space. When this occurs, the confidence intervals considered in this paper are appropriately truncated so that they only contain values in the parameter space. The coverage probabilities, of course, are unaffected by this truncation. One may also note that $E_{\rho_T}\left[U(\boldsymbol{Q}) - L(\boldsymbol{Q})\right] \leq 1$ since the limits of integration in (9) are from 0 to 1.

3 Example

Data were obtained on one hundred and seventy one yearling bulls from a Red Angus seed stock herd in Montana (Evans et al. (1995)). A trait of interest was the loineye (i.e., ribeye) muscle area measured in square inches. Ultrasound techniques were used to procure these

measurements. The measurement was taken on the dorso-ventral line between the 12th and 13th ribs on the left side of each animal. Figure 3 displays a histogram of the loineye data. The fixed effect was age of dam which had been originally recorded as belonging to one of eight categories: 2 years, 3 years, 4 years, 5-9 years, 10 years, 11 years, 12 years, and 13 or more years. Since there were only a few observations associated with dams greater than or equal to 10 years of age, our analysis used five categories for age of dam: 2 years, 3 years, 4 years, 5-9 years, and 10 or more years. The random effects are the animal's (additive) genetic effect and error.

The mixed linear model we consider is

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{10}$$

where $\boldsymbol{Y}$ is a 171×1 vector of observable random variables, $\boldsymbol{X}$ is a 171×5 incidence matrix, $\boldsymbol{\beta}$ is a 5×1 vector of unknown parameters, $\boldsymbol{Z} = \boldsymbol{I}_{171}$, and $\boldsymbol{u}$ and $\boldsymbol{e}$ are vectors of unobservable random variables of size 171×1.

The relationship matrix, denoted by $\boldsymbol{A}$, was determined using a recursive method given in Henderson (1976). It uses knowledge of the animal's sire, dam, and grandparents. Note that some animals are inbred so that it is possible that $Var(u_i) > \sigma_a^2$. For instance, it turns out that $Var(u_1) = 1.03125\sigma_a^2$.

The number of distinct eigenvalues of $\boldsymbol{H'ZAZ'H}$ is $d = 165$. Eigenvalues range in magnitude from $\Delta_1 = 0.56569$ to $\Delta_{165} = 8.65925$. Except for $\Delta_{61} = 0.67188$ having $r_{61} = 2$, all eigenvalues have a multiplicity of one. There are one hundred and sixty four possible versions of the pivotal quantity of the form (4) that can used to construct confidence intervals for ρ. In this paper we consider equal-tailed confidence intervals where $\alpha_1 = \alpha_2 = 0.05$.

Figure 4 depicts the expected lengths for selected confidence intervals as a function of ρ_T. The numerical label on each curve indicates the value of k. Recall that k is the number of quadratic forms in the denominator of the pivotal quantity (4). Over the entire parameter space, $k = 150$ and 155 correspond to those confidence intervals having relatively short expected lengths. Although not shown in Figure 4, the confidence intervals corresponding to values of k between 150 and 155 also have short expected lengths. In fact, Figure 4 suggests for a fixed value of ρ_T the expected length is minimum when k is between 150 and 155. The value of k that corresponds to the shortest expected length depends on the value ρ_T. In other words, for this application of the pivotal quantity technique, there does not exist an interval having minimum expected length across the entire parameter space.

The actual confidence intervals were computed using the one hundred and seventy one loineye measurements. In many cases, inverting the pivotal quantity in (4) results in confidence intervals whose endpoints fall outside of the parameter space. That is, inverting $F_{\alpha_1} \leq \frac{\sum_{i=k+1}^{d} \frac{Q_i}{1+\rho(\Delta_i-1)} / \sum_{i=k+1}^{d} r_i}{\sum_{i=1}^{k} \frac{Q_i}{1+\rho(\Delta_i-1)} / \sum_{i=1}^{k} r_i} \leq F_{1-\alpha_2}$ may result in $L < 0$ and $U \geq 1$ where L and U are illustrated in Figure 2. Of the one hundred and sixty four confidence intervals, the shortest interval is $(0.00, 0.42)$. This confidence interval corresponds to the pivotal quantity having $k = 153$. It should be noted that employing expected length as a criterion for good

confidence intervals does not guarantee the selected interval from a single realization will have minimal length.

4 Summary

Confidence intervals for ρ provide information concerning the likely values of the heritability of a specific trait under study. The endpoints of the intervals are a function of the data through the quadratic forms $Q_1, ..., Q_d$. The number of quadratic forms depend on the mixed linear model employed as well as the relationships among the animals. For the data set in Section 3, $d = 164$.

In general, there are many versions of the pivotal quantity that may be used to obtain confidence intervals for ρ. In this paper we have examined expected length properties with the intention of providing guidance to users for selecting a member of this family in any particular application. It seems reasonable to select the confidence interval that has relatively short expected lengths across the entire parmeter space. Expected length computations depend on the structure of the mixed linear model and not on the actual observations. The investigator may compute expected lengths without looking at the data and then select the pivotal quantities that yield short expected lengths. The actual value of k that corresponds to the shortest expected length depends on ρ_T, Δ_i, and r_i, $i = 1, ..., d$.

For the data set in this paper, the pivotal quantities having $k = 150$ through $k = 155$ result in intervals having short expected lengths. That is, we have narrowed the search for "good" confidence intervals from one hundred and sixty four to just a few. If the investigator believes ρ_T is small, Figure 4 indicates $k = 155$ is a good choice. If ρ_T is large, $k = 150$ is a good choice. For intermediate values of ρ_T, values of k between 150 and 155 are adequate and produce similar results.

The data set considered above exemplifies the fact that in many cases there does not exist a uniformly minimum expected length confidence interval. The computed expected lengths depend heavily on the true value of the parameter. This prompts one to consider additional criteria such as minimax or minimizing the average expected length, where the averaging is over the possible parameter values having preassigned weights. Similarly, Bayesian analysis could be employed by assuming an appropriate prior distribution for ρ.

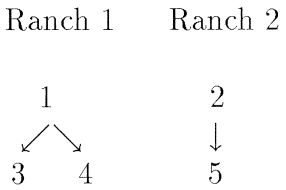

Figure 1: Pedigree Structure

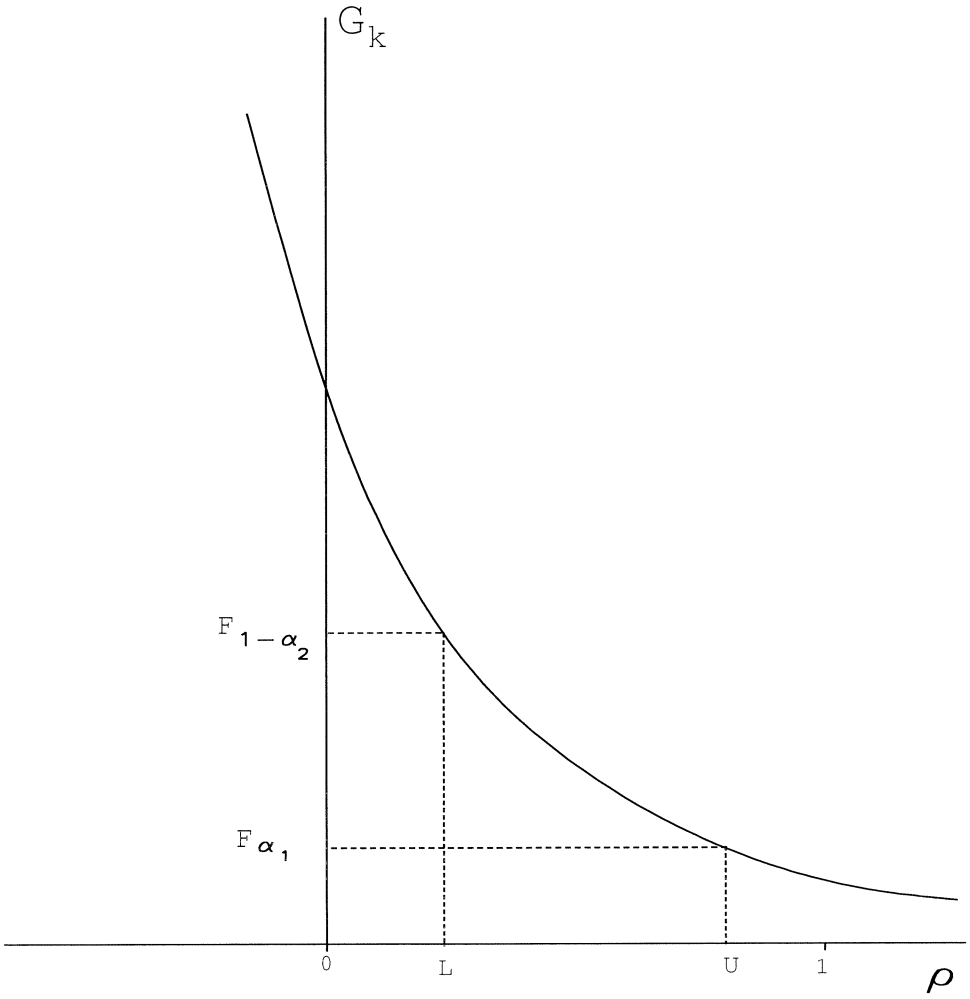

Figure 2: Inverting the Pivotal Quantity

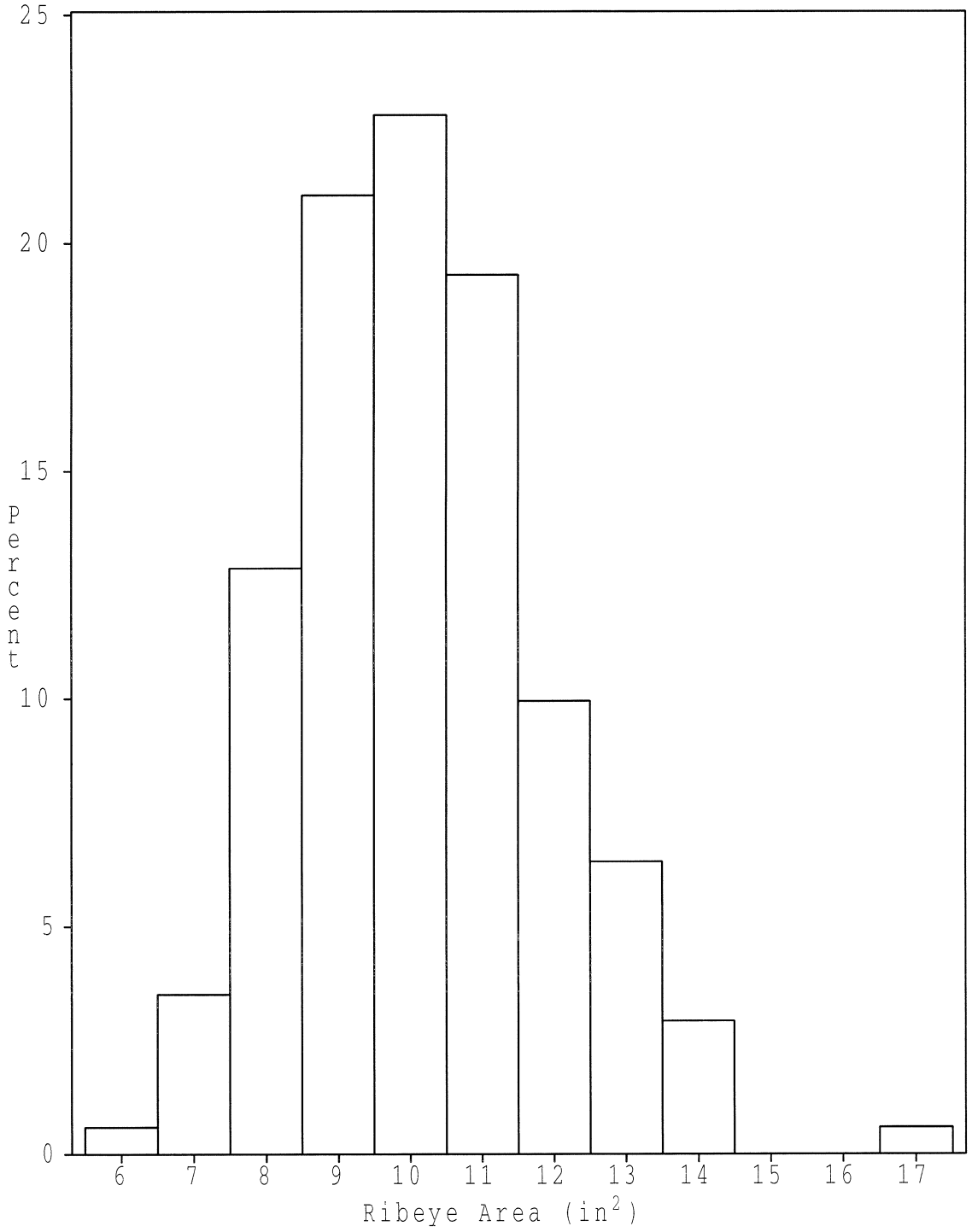

Figure 3: Histogram of Loineye Data

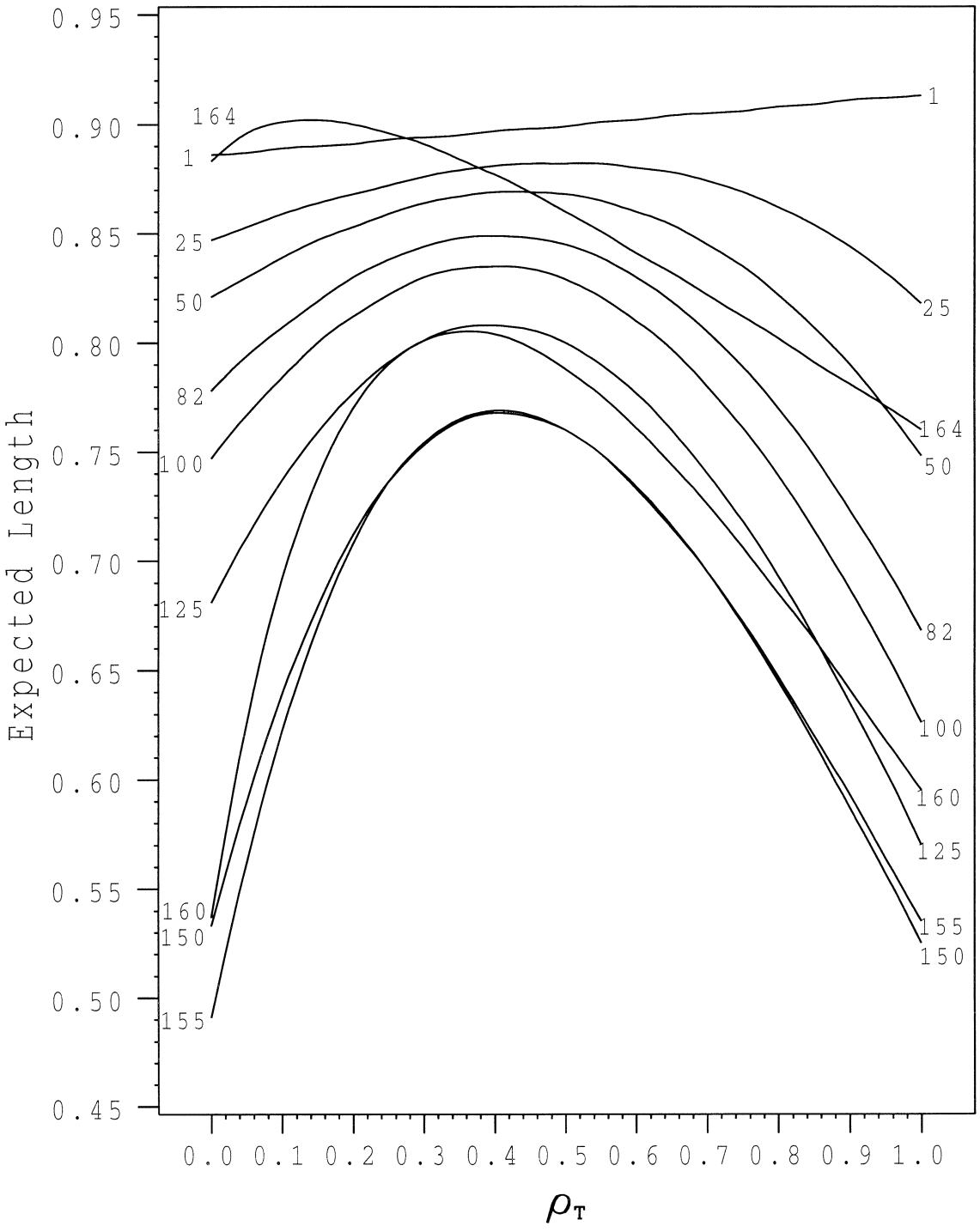

Figure 4: Expected Length of 90% Confidence Intervals for Loineye Data

References

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29,** 323–333.

Evans, J. L., Golden, B. L., Bailey, D. R. C., Gilbert, R. P. and Green, R. D. (1995). Genetic parameter estimates of ultrasound measures of backfat thickness, loineye muscle area, and gray shading score in red angus cattle. *Proceedings, Western Section, American Society of Animal Science* **46,** 202–204.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics.* Essex, England: Longman Scientific & Technical.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32,** 69–83.

Imhoff, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48,** 419–426.

LaMotte, L. R. and McWhorter, A., Jr. (1978). An exact test for the presence of random walk coefficients in a linear regression model. *Journal of the American Statistical Association* **73,** 816–820.

LaMotte, L. R., McWhorter, A., Jr. and Prasad, R. A. (1988). Confidence intervals and tests on the variance ratio in random models with two variance components. *Communications in Statistics - Theory and Methods* **17,** 1135–1164.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association* **56,** 549–567.