BAYES RISK ANALYSIS OF REGIONAL REGRESSION ESTIMATES OF FLOODS

by

William Arledge Metler

A Thesis Submitted to the Faculty of the

DEPARTMENT OF SYSTEMS AND INDUSTRIAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE
WITH A MAJOR IN SYSTEMS ENGINEERING

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 7 2

STATEMENT BY AUTHOR

SIGNED: William A. Mitler

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Lucien Duckstein
LUCIEN DUCKSTEIN
Professor of Systems and
Industrial Engineering

1 dec. 1972
Date

ACKNOWLEDGMENTS

This study could not have proceeded as well as it has without the assistance and understanding of many people. The author is indebted to Dr. Lucien Duckstein whose assistance, advice, and foresight provided the impetus and leadership in the organization of this thesis. The author is also highly indebted to Dr. Chester C. Kisiel as he was responsible for this student having the opportunity to work with the USGS. His patience and experience had a significant impact on the progress of this thesis.

The local USGS office was quite helpful in rendering reference material. At the local office, Dave Click and Lucela Siberts were quick to respond to all administrative requests. Byron Aldridge, in the initial investigation, was helpful in familiarizing the author with current USGS techniques. Mr. D. M. Thomas readily supplied the necessary data for the computer program. Mr. Dave Dawdy deserves special mention as he was instrumental in defining the interests of the USGS.

The author wishes to express the greatest gratitude to Dr. Don R. Davis. His patience in explaining concepts and his practical experience were invaluable. His criticisms and comments were always pertinent and were regarded with great confidence.

The author would like to thank Mrs. Jessie Cude for her patience and professionalism in typing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ABSTRACT

This thesis defines a methodology for the evaluation of the worth of streamflow data using a Bayes risk approach. Using regional streamflow data in a regression analysis, the Bayes risk can be computed by considering the probability of the error in using the regionalized estimates of bridge or culvert design parameters. Cost curves for over- and underestimation of the design parameter can be generated based on the error of the estimate. The Bayes risk can then be computed by integrating the probability of estimation error over the cost curves. The methodology may then be used to analyze the regional data collection effort by considering the worth of data for a record site relative to the other sites contributing to the regression equations.

The methodology is illustrated by using a set of actual streamflow data from Missouri. The cost curves for over- and underestimation of the streamflow design parameter for bridges and culverts are hypothesized so that the Bayes risk might be computed and the results of the analysis discussed. The results are discussed by demonstrating small sample bias that is introduced into the estimate of the design parameter for the construction of bridges and culverts. The conclusions are that the small sample bias in the estimation of large floods can be substantial and that the Bayes risk methodology can evaluate the relative worth of data when the data are used in regionalization.

CHAPTER 1

INTRODUCTION

Given unlimited funds and resources, an engineer could design and build a culvert or bridge which might last forever. However, because both funds and resources are limited, the engineer must trade off the expenditure of funds and the use of resources against the life of the culvert or bridge. The engineer, then, seeks to optimize the use of funds and resources within the constraint of the life of the culvert or bridge. At a given construction site, if the design criteria were set in terms of the 25-year flood and if the true discharge of the 25-year flood for the site were known, then the construction specifications for the culvert or bridge could be computed from a study of the characteristics of the materials and the hydraulic behavior of the flows passing through the culvert or under the bridge. However, the true discharge of the design flood is <u>not</u> known but can only be estimated. Thus there are two problems to be solved in the design of culverts or bridges. The first problem is the estimation of the flow of the design flood. The second problem is the hydraulic design of the culvert or bridge. This thesis will be concerned only with a portion of the estimation problem. The issues posed above are also relevant to many other engineering design situations.

In order to estimate the design flood at a construction site, one might establish a data collection effort at the site, and based upon this data, one might by some technique compute an estimate of the design flood. However, here again, time, money, and resources are constraints. Currently, it is deemed infeasible to establish a data collection effort at a particular construction site each and every time an estimate of a design flood is required. It would seem feasible, however, to establish a data collection network for a region whereby one could predict the design flood from the data which have already been collected. Such a data collection network for a region has been established by the U. S. Geological Survey (USGS). Regional data collection leads to regionalization, and one method of generalization relies on regression analysis. A streamflow characteristic, e.g., the 25-year flood, is regressed upon basin characteristics, e.g., basin area, channel slope, local precipitation, etc. Given the basin characteristics of an ungaged site, i.e., the design site, the resulting regression equation then can be used to predict that streamflow characteristic. However, the USGS is constrained by time, money, and resources. The USGS, then, is concerned with the trade-off of the costs of collection versus the payoffs of collection.

This thesis will define a methodology for quantifying the payoff of collecting data. By collecting more streamflow data the uncertainty (probable error) in the historical estimate of the design flood for a data site may be reduced. That is, by lengthening the historical record at a site, more information about the hydrologic process for the site is generated. In particular, for a record of annual peak flows, the information content of the record relates to the moments of the probability density function (pdf) of the peak flows. The moments of the probability density function of the peak flows must be estimated from statistics, e.g., the mean and the variance. Because these statistics are estimates, there is uncertainty about their representing the true values of the parameters of the pdf. Fisher (1949) and Matalas (1967) suggest that the information about a parameter or streamflow characteristic may be given by the reciprocal of the variance of the parameter or characteristic. The variance approach is one way to quantify our uncertainty about a parameter.

However, the variance approach fails to consider the risks of error in the estimate. Consider, for example, two estimates with equal variances. One estimate is used to design a bridge for a secondary highway. The other estimate is used to design a larger bridge for an access road to a hospital. Although the information about the estimates are equal by the variance approach, there is greater loss associated with the error in the estimate for the hospital bridge than for the other bridge. Both the social and economic costs for washout or damage to the hospital bridge are higher than for the other bridge. On the other hand, the structural costs for overbuilding the hospital bridge may be higher than for the other bridge. For this case, then, the variance approach would be inadequate in attempting to measure the accuracy of an estimate (Davis, Kisiel, and Duckstein 1972).

This thesis assumes that a decision about the optimal design flood has already been made. The decision maker is uncertain about the true state of nature in regard to the life of the culvert. Green and

Tull (1970, p. 17) define uncertainty as "a state of doubt . . . with respect to outcomes, given a particular course of action." Even though his decision is optimal (expedient), there is a risk (liability) associated with the implementation of this decision. Klausner (1969, p. 184) defines risk as "the consequential effect of possible uncertain outcomes." For example, management has specified that the design flood for a culvert is the 25-year flood. The truly optimal design flood may not be the 25-year flood, but the decision makers have chosen the 25-year flood according to their knowledge of the situation. In making such a decision, management is implying that it is willing to accept the risks of this decision. It is not the subject of this study to determine how this decision was judged optimal with respect to the underlying risk. In this context, a precise definition of the design risk (as opposed to the later defined risk in estimation) is not possible, but a vague notion of a weighted trade-off of several factors is implied. If the true discharge of the 25-year flood were known, then the culvert could be constructed and would be considered optimal according to the management's decision. However, the 25-year flood must be estimated by some technique. This design estimate serves as the optimal design flood. There is uncertainty about the appropriateness of the estimate to act as the true design flood which has been specified as being optimal. Hence, if there is an error in the estimate, then there is a loss associated with the use of this incorrect estimate of the optimal design flood. In the decision phase, there is an uncertainty about the optimality of a design parameter, whereas in the estimation phase there is an uncertainty about the estimate's being the true value of the parameter. Our knowledge or uncertainty about the true value is conditioned upon the estimate (i.e., how we use the data) and the accuracy (variance) of this estimate. The loss due to error in the estimate given by the optimal decision and the probability of all possible errors are combined to give the Bayes risk. The Bayes risk is the expected loss of using an estimate as the optimal design parameter. Bayes risk is also called the terminal expected opportunity loss by Raiffa and Schlaifer (1961). Thus, the Bayes risk will refer to the expected consequential effects of using the computed estimate of a design flood which has already been chosen to be optimal.

The first section of this thesis defines a methodology for evaluating the Bayes risk in using some estimate of a bridge or culvert design parameter. This design parameter is an estimate of a streamflow characteristic as computed from the regional data or possibly directly from the historical data. The true value of the design parameter, e.g., the 25-year flood, is not known, but our knowledge of it or uncertainty about it can be described by a probability density function (pdf) and is called the error pdf. Losses caused by error either in underestimation or overestimation of the true value can be expressed in cost functions. For this study, losses caused by underestimation will be expressed in an underdesign loss curve and losses caused by overestimation will be expressed in an overdesign loss curve. The combination of the error pdf with the underdesign loss curve and the overdesign loss curve will yield an expected loss in using a given estimate as the design parameter for the construction of a bridge or culvert. This report is not concerned with the exact formulation of the cost (loss) curves but will hypothesize the cost curves so that the methodology may be illustrated. Since the methodology combines both our uncertainty about an estimate and the costs due to the uncertainty in the estimate, then the Bayes risk is a measure of the worth of data or of the information content of a data set.

The remaining chapters of the report are concerned with defining the error pdf of the estimate to be used in computing the Bayes risk. Chapter 3 is concerned with the estimation of the design parameter by a regional regression equation. A review of regression is presented, and the error pdf for the regression estimate is approximated in this report by a normal distribution. The mean and variance of this distribution are derived. Thus the Bayes risk for the regression estimate may be computed by the methodology if the hypothetical cost curves are accepted.

The next chapters are concerned with two types of historical estimates. The historical estimate differs from the regression estimate in that the historical estimate is computed directly from the data record for a particular site whereas the regression estimate is inferred from all the data for a region. The historical estimate is important because it serves as input to the regression analysis which is used for computing the regression estimate.

The first type of historical estimate to be discussed is what will be called the normal-estimate. For this report a log normal distribution of peak annual discharges is assumed. Thus, for a particular recurrence interval, e.g., the design parameter of the 25-year flood, the discharge can be estimated from the log normal pdf. Because the estimate of the discharge is a statistic, there is uncertainty about its being the true design parameter. This uncertainty can be expressed as a pdf which is a function of the normal-estimate and the variance of that estimate.

The second type of historical estimate is a refinement of the first type. Because most historical records are comparatively short, there is a small sample bias introduced into the computation of the estimate of the design parameter by the normal method. If a log normal distribution of peak annual discharges is assumed, then this small sample bias is removed if the t-distribution is used to compute the estimate of the design parameter. The error pdf for this t-estimate is assumed to be approximately normal with a mean given by the computed estimate and variance as derived in this paper. In this chapter the mean and variance of the error distribution of the t-estimate, because they fail to consider the effects of small sample size, will <u>always</u> be less than the results given by use of the Student's t-distribution.

The final section presents an example of the Bayes risk methodology. Because of the numerical problems in the example, a computer program was used to compute the Bayes risk for thirty data sites within a region. Four different iterations of the methodology were performed on the thirty sites. The first iteration computes the Bayes risk for each of the thirty sites based on the normal-estimate of the design parameter. The second iteration computes the Bayes risk for each of the thirty sites based based on the t-estimate of the design parameter. These two sets of Bayes risks are contrasted to demonstrate the significance of the small sample bias. The third iteration computes the Bayes risk for each of the thirty sites based on the regression estimate of the design parameter. This iteration is performed in two parts: the first part uses the normal-estimates as inputs to the regression, and the second part uses the t-estimates as inputs to the regression. Here again the results of the two sets of Bayes risk, for the normal-inputs and the t-inputs, demonstrate the significance of the small sample bias. The fourth iteration computes the Bayes risk for each of the thirty sites with each site in turn being excluded from the regression with the t-estimates serving as inputs. The set of Bayes risks with all sites included in the regression is then contrasted to the set of Bayes risks with each site excluded from the regression. The contrast identifies dramatically that site data set which is considered to have the highest Bayes risk. This data set is not necessarily the data set yielding the estimate with the greatest variance. Lastly, the thirty sites can be ranked according to increasing Bayes risk, and hence information content can be judged by considering relative risk liability.

The Bayes risk methodology, then, is considered to be an extension of the variance approach. Information content which was based on the reciprocal of the variance can now be measured by the risk of error in the use of the information. The next section defines a procedure for evaluating the worth of data by a Bayes risk approach.

CHAPTER 2

THE BAYES RISK METHODOLOGY

This chapter will define a methodology for the evaluation of the worth of data by a Bayes risk approach. We will first derive the error pdf for the estimate of the design parameter; then, we will define the loss curves. The loss due to error in the estimate will then be integrated over the error pdf to give the Bayes risk in the use of the estimate. In this discussion, the reader will be aided by Figure 1.

Let us assume that we have an estimate of a construction parameter for the design of a bridge or culvert. Specifically, this parameter is the discharge of the design flood. This estimate can be computed from streamflow data by several methods, some of which will be discussed later in this report. But for now let us assume that the peak annual discharges belong to a logarithmic transform of a population, e.g., log Pearson III or log normal. Unless otherwise stated, any reference to a discharge hereafter will be to the logarithm of that discharge.

Let the estimate of the design parameter be Q (the antilog of which would give us the flow). Suppose now that we compute a large number of estimates under exactly the same conditions. By the central limit theorem the distribution of these estimates will be a normal distribution as the number of replications tends to infinity. The square root of the variance of this distribution is the standard error of estimate. Our knowledge about the true value of the design flow QTRUE is

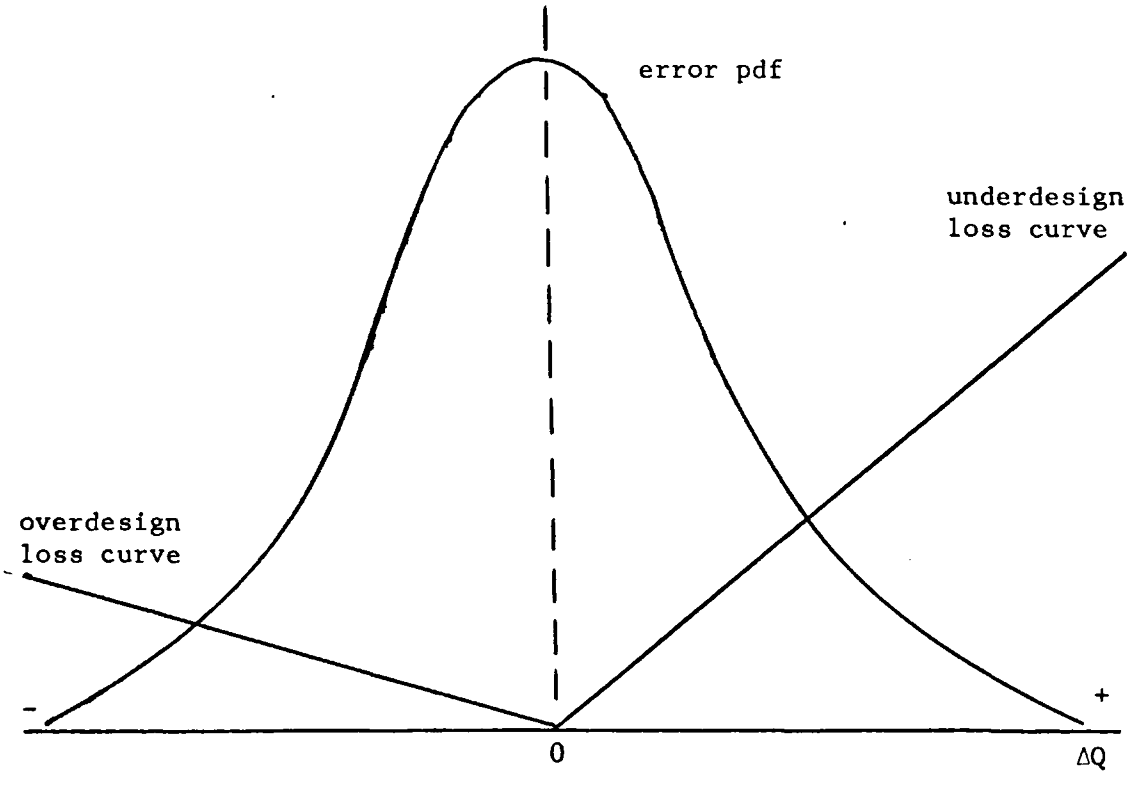

Figure 1. The error pdf and loss curves for errors ΔQ in flood estimates.

a random variable KQ which is distributed normally as described above. However, if we only have one estimate Q of QTRUE, the variance of this estimate acts as an estimator for the variance of the distribution of the estimates (Aitchison 1970). That is, our knowledge (according to the data) of the true discharge QTRUE is a random variable whose pdf is normal with mean Q and variance Var(KQ). In mathematical notation

$$KQ \sim N(Q, Var(KQ)). \tag{2.1}$$

We are quantifying our knowledge of QTRUE in (2.1) or approximating the distribution of the estimates by replication of experiments as described above. The error pdf can now be shown as a transform of (2.1). If Q = QTRUE then there is no error in estimation. Hence the error pdf is

$$ERR \sim N(0,1)$$

where (2.1) has been standardized. Hence, to compute the error in a specified probability interval, a transformation from the standard normal must be made back to (2.1). Throughout the remainder of this report, the error pdf will be synonymous with the knowledge pdf (2.1) where it is understood that a transformation is necessary to compute the error given the probability of error or vice-versa.

Next, the loss curves will be hypothesized so that the loss of a specified error may be computed and multiplied by its respective probability. As discussed in the Introduction, we are concerned with two types of cost curves; the first one is the underdesign loss curve. Assume QTRUE lies above Q and we design the culvert or bridge based on Q.

Thus, we have underestimated the true discharge and hence must suffer damage and social costs above those costs which we are billing to suffer with the design based on Q = QTRUE. Assume that this loss monotonically increases as the error in underestimation increases (Aitchison 1970). Define the loss (due to underestimation) as

$$L(\Delta Q) = a\Delta Q \tag{2.2}$$

where

$$\Delta Q = QTRUE - Q \tag{2.3}$$

for QTRUE $\geq$ Q and $L(\Delta Q) = 0$ for QTRUE < Q. In (2.2) a is a loss scaling factor peculiar to the construction site and the estimate (explained later in this chapter).

The other type of cost curve is the overdesign loss curve. Assume QTRUE lies below Q and we design the culvert or bridge based on Q. Thus we have overestimated the true discharge and hence must suffer an overdesign loss. That is, a smaller culvert or bridge could have been designed based on the given design parameter. Assume that the cost monotonically increases as the error of overestimation increases. Define the overdesign loss as

$$OL(\Delta Q) = b\Delta Q \tag{2.4}$$

where

$$\Delta Q = Q - QTRUE \tag{2.5}$$

for Q $\geq$ QTRUE and $OL(\Delta Q) = 0$ for Q < QTRUE. In (2.4), b is an overdesign loss scaling factor peculiar to the construction site and the estimate. In both (2.2) and (2.4) it is assumed that the loss is a function of the error in estimation and a loss factor peculiar to each construction site. The results of such formulations of loss will be seen in the example chapter.

Both of the formulations (2.2) and (2.4) can be viewed in the following manner. Consider that we have an unlimited number of duplicates of site k. Furthermore, we design and build a culvert for site k based on Q(k). Now assume that QTRUE(k) is the same for all duplicates of site k and that after construction we are told QTRUE(k). Thus, we have over- or underdesigned each of the culverts by the same amount. Now assume that all sites k have the same a and b loss coefficients. During the design life d of the culvert, we are able to collect and tabulate every kind of loss incurred due to an inaccurate design estimate. For this particular error then, we are able to determine an average loss due to either over- or underestimation relative to the case when QTRUE = Q. Now we repeat this process for all values of error, and thus we are able to define the loss curves formulated in (2.2) and (2.4). The formulation of actual OL and L curves is a sizable problem in itself, and hence this report will use the OL and L curves of the form of (2.2) and (2.4). For this report the quantities a and b are selected arbitrarily but in such a way that the numerical values for the Bayes risks are manageable in the computerized example. We are now ready to combine the error pdf with the loss curves.

The Bayes risk is the sum of the expected loss due to underdesign and the expected loss due to overdesign. The expected loss due to underdesign is given by

$$E(L) = \int_0^\infty L(KQ)N(KQ|Q,Var(KQ))dKQ \ . \tag{2.6}$$

The expected loss due to overdesign is given by

$$E(OL) = \int_{-\infty}^0 OL(KQ)N(KQ|Q,Var(KQ))dKQ \ . \tag{2.7}$$

In (2.6) and (2.7) L(KQ) is defined by (2.2), OL(KQ) is defined by (2.4), and N(KQ|Q,Var(KQ)) is the normalized (2.1). Thus the Bayes risk is

$$E(R) = E(L) + E(OL) \tag{2.8}$$

Equations (2.6), (2.7), and (2.8) thus define the Bayes risk in using Q as an estimate of the design parameter. The loss curves have been defined in (2.2) and (2.4). We can now proceed to define the parameters of the error pdf. The first method for defining Q, Var(KQ), and the error pdf will be the method which employs the streamflow data on a regional scale. This regionalization of data employs regression analysis.

CHAPTER 3

WORTH OF DATA FOR REGRESSION

A Review of Regression

The U. S. Geological Survey has adopted the use of regression analysis as the form for the regionalization of streamflow data (Thomas and Benson 1970). A brief review of regression is given here so that the reader will understand its use in the methodology and will become familiar with the notation used in this report. For further information about regression the reader is referred to the references.

In regression analysis a set of independent variables are linearly related to a dependent variable. In the regionalization of streamflow data, the set of independent variables is composed of selected topographical and climatic characteristics of a drainage basin, and the dependent variable is a streamflow characteristic, e.g., the 25-year flood. A hydrological region might be defined as a geographical area with similar basin and streamflow characteristics (Thomas and Benson 1970).

Consider a region with n gaged sites with each site being described by p topographical and climatic characteristics. A particular streamflow characteristic, then, might be described by the equation

$$D_d = \beta_0 C_1^{\beta_1}, C_2^{\beta_2}, \ldots, C_p^{\beta_p} \gamma \tag{3.1}$$

where C_i, $i = 1, \ldots, p$ are basin characteristics such as drainage area, forest cover, precipitation, elevation, etc.; D_d is the estimate of the design discharge which must be estimated from the peak annual discharge record; and γ and β_i, $i = 1,\ldots,p$ are parameters of the regression equation. By taking logs, (3.1) becomes

$$QTRUE = \beta_0 + \beta_1 \log C_1 + \beta_2 \log C_2 + \ldots + \beta_p \log C_p + \eta \tag{3.2}$$

and this then is the multiple linear regression equation. In (3.2) $\eta = \log\gamma$ represents a random component which is the departure of the dependent and independent variables from a direct linear relationship. For the n sites it is assumed that the η_i are independently and identically distributed random variables with mean 0 and common variance

$$\begin{aligned} E(\eta_i) &= 0 \qquad 1 = 1,\ldots,n \\ E(\eta_i \eta_j) &= \sigma^2 \quad i = j = 1,\ldots,n \\ &= 0 \qquad i \neq j \end{aligned} \tag{3.3}$$

In actual practice, however, QTRUE, σ^2, and β_i are not known and must be estimated. For site k the estimation of QTRUE(k) by Q(k) will be discussed later. For the present let us assume that QTRUE(k) = Q(k) and then proceed to estimate the β_i and σ^2.

Let the n observations of the p independent variables be given by the matrix

$$\overline{C} = \begin{bmatrix} 1 & \log c_{11} & \log c_{12} \ldots \log c_{1p} \\ 1 & \log c_{21} & \log c_{22} \ldots \log c_{2p} \\ \ldots & & \\ 1 & \log c_{n1} & \log c_{n2} \ldots \log c_{np} \end{bmatrix} \qquad (3.4)$$

and the historical observations of the dependent variable by the vector

$$\overline{QH} = [QH_{(1)}\ QH_{(2)} \ldots QH_{(n)}]^T \qquad (3.5)$$

where T represents the transpose and QH(k) is an estimate of the QTRUE(k) computed directly from the historical record. The vectors

$$\overline{\beta} = [\beta_0\ \beta_1 \ldots \beta_p]^T \qquad (3.6)$$

and

$$\overline{\eta} = [\eta_1\ \eta_2 \ldots \eta_p]^T \qquad (3.7)$$

give the regression coefficients and the random components, respectively. Let the estimates of the population regression coefficients β_i be

$$\overline{b} = [b_0\ b_1 \ldots b_p]^T \qquad (3.8)$$

Thus, (3.2) for n sites becomes

$$QH_{(1)} = b_0 + b_1 \log c_{11} + b_2 \log c_{12} + \ldots + b_p \log c_{1p} + \epsilon_1$$

$$QH_{(2)} = b_0 + b_1 \log c_{21} + b_2 \log c_{22} + \ldots + b_p \log c_{2p} + \epsilon_2$$

$$\ldots$$

$$QH_{(n)} = b_0 + b_1 \log c_{n1} + b_2 \log c_{n2} + \ldots + b_p \log c_{np} + \epsilon_n \qquad (3.9)$$

where ϵ_k is an estimate of the kth random component and is called the residual for site k. The regression estimate QR(k) is then

$$QR(k) = b_0 + b_1 \log c_{k1} + b_2 \log c_{k2} + \ldots + b_p \log c_{kp} . \quad (3.10)$$

The method of least squares seeks to minimize the sum of squares of the residuals

$$\sum_i (QH(i) - QR(i))^2 = \sum_i \epsilon_i^2 . \quad (3.11)$$

That is, the estimates of the regression coefficients b_i are computed so that the sum of squares of the differences between the observed flow and the regression prediction is minimum. Using calculus (Smillie 1966) it can be shown that (3.11) is minimum when the matrix equation

$$\overline{h} = \overline{A}\,\overline{b} \quad (3.12)$$

is solved for $\overline{b}$. In (3.12) $\overline{b}$ is given by (3.8),

$$\overline{A} = \begin{bmatrix} n & \Sigma\log c_{i1} & \ldots & \Sigma\log c_{ip} \\ \Sigma\log c_{i1} & \Sigma\log c_{i1}^2 & & \Sigma\log c_{i1}\log c_{ip} \\ \ldots & & & \\ \Sigma\log c_{ip} & \Sigma\log c_{i1}\log c_{ip} & \ldots & \Sigma\log c_{ip}^2 \end{bmatrix} \quad (3.13)$$

and

$$\overline{h} = [\Sigma QH_i \ \ \Sigma QH_i \log c_{i1} \ \ldots \ \Sigma QH_i \log c_{ip}]^T . \quad (3.14)$$

The solution is

$$\overline{b} = \overline{A}^{-1}\overline{h} \quad (3.15)$$

where $\bar{A}^{-1}$ is nonsingular if any one set of observations is not a linear combination of another. Let

$$\bar{D} = \bar{A}^{-1} = \begin{bmatrix} d_{00} & d_{01} & d_{02} & \cdots & d_{0p} \\ d_{10} & d_{11} & d_{12} & \cdots & d_{1p} \\ \cdots & & & & \\ d_{n0} & d_{n1} & d_{n2} & \cdots & d_{np} \end{bmatrix} \tag{3.16}$$

where $\bar{D}$ is known as the variance-covariance matrix. An alternate notation used in some texts is (e.g., Draper and Smith 1966)

$$\bar{b} = (\bar{C}^T\bar{C})^{-1} \bar{C}^T \overline{QH} \tag{3.17}$$

where it will be seen that $(\bar{C}^T\bar{C})^{-1} = \bar{D}$ and $\bar{C}^T \overline{QH} = \bar{h}$. An unbiased estimate of the residual variance σ^2 is

$$s^2 = \frac{1}{n-p-1} \sum_i \epsilon_i^{\ 2} \tag{3.18}$$

Thus far no assumptions have been made about the distributions of the dependent and independent variables. If it is assumed that the independent variables belong to a multivariate normal population, then the dependent variable belongs to a normal population (Morrison 1967). With this assumption the means of the conditional distributions of the dependent variables lie on a straight line. The linearity assumption can be tested by statistically examining the residuals. In order to test statistically the residuals and compute confidence intervals for the b_i, it must be assumed that the residuals are identically and independently distributed random variables. That is, the residuals are homoscedastic (have common variance) and belong to a normal population with mean zero (3.3). The Durbin-Watson statistic (Smillie 1966) is one test of autocorrelation of residuals (indicative of nonlinearity). The multiple correlation coefficient, $0 \leq R^2 \leq 1$, is one measure of the applicability of the regression model. A high R^2 indicates that a large proportion of the variance of the dependent variable may be removed by knowledge of the independent variables. The use of the regression model for streamflow prediction is not the primary subject of this report, and hence the applicability and shortcomings of the model will not be discussed in greater detail. The interested reader may refer to several USGS reports and texts [e.g., Thomas and Benson (1970) and Draper and Smith (1966)] for further information about the use of regression analysis. The next section applies the methodology of Bayes risk to the use of the regression estimate.

Application to the Regression Estimate

This section defines the error pdf for the methodology outlined previously for the evaluation of the Bayes risk. The error pdf of the regression estimate must be defined in order to evaluate the Bayes risk in the use of this estimate as a design parameter.

As discussed in the previous section, the prediction of a streamflow design parameter for site k may be given by

$$QR(k) = b_0 + b_1 \log c_{k1} + b_2 \log c_{k2} + \ldots + b_p \log c_{kp}$$

which is the mean of a conditional distribution. It will be hypothesized here that the error pdf is given by the normal distribution. This hypothesis follows from the normality assumption of the residuals. Let the mean of the uncertainty distribution for site k, then, be given by the regression estimate QR(k). The variance of this distribution will be derived from the use of confidence intervals.

Several texts (Smillie 1966, Draper and Smith 1966) give the $100(1-\alpha)$ percent confidence limits for an individual prediction of a dependent variable as

$$QR(k) \pm t_{\alpha/2} \left(s^2 \left(1 + \sum_{i,j=0}^{p} d_{ij} \log c_{ki} \log c_{kj}\right)\right)^{\frac{1}{2}} \tag{3.19}$$

where $t_{\alpha/2}$ comes from tables of the Student's t at the α probability level, s^2 is defined by (3.18), and d_{ij} by (3.16). In order to establish the variance in terms of confidence limits, an interpretation of confidence intervals is given below.

A confidence interval is interpreted as follows. Consider the regression estimate QR(k) which is an estimate of the conditional mean. Of necessity QR(k) must act as the prediction of any particular value of the random variable expressed by the conditional distribution at a point on the regression line. Consider then that we are interested in finding the 95% confidence interval for the true design parameter QTRUE(k) given a specific set of values of the independent variables. Define the probability interval to be

$$P(A(QR(k)) \leq QTRUE(k) \leq B(QR(k))) = 1 - \alpha \tag{3.20}$$

where A(QR(k)) is the lower limit given by (3.19) and B(QR(k)) is the upper limit. This is a probability statement about the interval, the interval itself being the random variable. The probability that the random variable, the interval, will contain the true value between the limits given in (3.20) is .95 for $\alpha = .05$. Once the limits become fixed by computation of QR(k) then the probability interval becomes a confidence interval. No longer can a probability statement be made since the random variable has assumed a specific value. The probability is either 1 or 0 that the fixed interval contains QTRUE(k) depending on whether the interval does or does not contain it. If from (3.20) the confidence intervals were calculated for each QR(k) computed in identical situations (experiments), then 95% of the confidence intervals resulting from such calculations would contain QTRUE(k). We cannot state that the probability of the computed interval containing QTRUE(k) is .95, although we may say it for practical purposes (Draper and Smith, 1966, p. 122). Hence, the use of confidence intervals for QTRUE(k) does not define the distribution of the error since we are not interested in the probabilities of QTRUE(k) lying within a certain rather wide interval but rather we should be interested in defining the probability that, for purposes of numerical integration, a very small specific interval does contain QTRUE(k).

It is shown in many texts (Draper and Smith 1966) that the confidence limits for an individual prediction widen as the conditional mean draws away from the mean of the dependent variable. This tendency should be reflected in the error pdf. Notice in (3.19) that the interval is a function not only of a value drawn from tables of the Student's t, but also a function of the residual variance s^2, the independent variables (c_{ij}'s), and the variance-covariance matrix (d_{ij}'s). Hence, define the error pdf of the regression estimate as

$$KQR(k) \sim N(QR(k),\ Var(KQR(k))) \tag{3.21}$$

where

$$Var(KQR(k)) = s^2 \left(1 + \sum_{i,j=0}^{p} d_{ij} \log c_{ki} \log c_{kj}\right) . \tag{3.22}$$

That is, Var(KQR(k)) now reflects the greater uncertainty of a conditional mean QR(k) which is further from the mean of the dependent variable. Hence, the error pdf reflects the conditional mean and shows the tendency of the confidence limits to widen.

The next chapter defines one method by which the inputs to the regression equation, the historical estimates, may be computed.

CHAPTER 4

THE BAYES RISK OF THE NORMAL-ESTIMATE

This chapter defines what will be called the normal-estimate of the design parameter. The normal-estimate is computed directly from the historical record. Later the error pdf of this estimate is defined so that the Bayes risk methodology may be applied to the normal-estimate. In Chapter 6 the Bayes risk of the normal-estimate will be contrasted with the Bayes risk of the t-estimate which is discussed in Chapter 5.

Consider a region of n gaged sites with each site being characterized by p variables. Each of the n gaged sites will have peak annual discharge records of differing lengths. Let the record length of site k be m(k). Assume that the annual peak flows for site k, k = 1, 2, ..., n, belong to a log normal distribution. The log normal distribution is not as general as the log Pearson III but may be used in some cases (Water Resources Council 1967). The analysis in this thesis is valid only for the log normal distribution. Hence, the logs of the flows will belong to a normal distribution. That is,

$$X \sim N(\mu_k, \alpha_k^{\,2}) \tag{4.1}$$

where X is the random variable expressing the log of the peak annual discharge (Q represents a design flow rather than just any flow), μ_k is the mean in logs of the normal pdf for site k, and $\sigma_k^{\,2}$ are unknown, they must be estimated from the data record. Let

$$\text{XBAR}(k) = \frac{\Sigma\, x\,(i,k)}{m(k)} \tag{4.2}$$

be the point estimate of μ_k and

$$\text{SHAT} = \left[\frac{\Sigma(x(i,k) - \text{XBAR}(k))^2}{m(k) - 1}\right]^{\frac{1}{2}} \tag{4.3}$$

be the point estimate of σ_k where x(i,k) is the log of the ith year peak annual discharge for site k. Hence, the design parameter, e.g., the 25-year flood, can be estimated by

$$\text{QHN}(k) = \text{XBAR}(k) + K(d)\ \text{SHAT}(k) \tag{4.4}$$

where QHN(k) is the historical normal-estimate for site k of the design flood, d is the recurrence interval of the design flood, and K(d) is the standard normal deviates for the $\frac{1}{d}$ exceedance probability which may be found in tables of the cumulative standard normal distribution. Equation (4.4) has been used by Nash and Amorocho (1966) and Hardison and Jennings (1972) to compute an estimate of the design flood.

For example, let XBAR = 5.85, SHAT^2 = .37, m(k) = 13, and d = 50 (K(50) = 2.054). Then

$$\text{QHN}(k)\,50 = 5.85 + (2.054)\ (.6)$$

$$\text{antilog QHN}(k)\,50 = 1210\ \text{ft}^3/\text{sec.}$$

Thus QHN(k) may be computed for each k = 1, 2, ..., n and this set of estimates along with the respective set of basin characteristics will serve as input to the regional regression analysis as outlined in Chapter 3.

Now let us proceed to the definition of the error pdf for the normal estimate.

Assume that a bridge or culvert is to be constructed at site k which happens to have m(k) years of record. Design will be based on the estimate QHN(k), but we run the risk of under- or overestimation. Thus to apply the Bayes risk methodology to evaluate the information content of the record for site k, we must define the error pdf and the cost curves.

Assume that the error pdf is approximately normal. Hardison and Jennings (1972) will demonstrate this in a Monte Carlo simulation. The mean of the pdf will be QHN(k). The variance of this pdf is the variance of the random variable KQHN(k), denoted Var(KQHN(k)). Nash and Amorocho (1966) and Hardison (1969) give

$$\mathrm{Var}(\mathrm{KQHN}(k)) = (\tfrac{1}{2}K(d)^2 + 1)\frac{\sigma^2}{m(k)} \tag{4.5}$$

where σ^2 can be approximated by SHAT^2. Actually (4.5) is a limiting case of a more general situation presented in the following chapter. Hence, the error pdf of the normal-estimate of the design flood as computed from the historical record is defined by

$$\mathrm{KQHN}(k) \sim N(\mathrm{QHN}(k), \mathrm{Var}(\mathrm{KQHN}(k))) \tag{4.6}$$

Both equations (4.4) and (4.5) assume that m(k) is of sufficient size to remove the bias of a small sample size. However, in many situations of streamflow estimation, m(k) is comparatively small so that the small sample size significantly affects the estimation techniques. The next chapter presents an estimation technique which removes the small sample bias. The estimate resulting from this formulation will be called the t-estimate.

CHAPTER 5

THE BAYES RISK OF THE t-ESTIMATE

In this chapter we will derive the t-estimate and an equation for evaluating the worth of data in using the t-estimate as the design parameter. The t-estimate is computed from the Student's t distribution and is derived below. Next, the error pdf for the t-estimate is derived so that the bias of the small sample size may be illustrated.

Again consider a region with n gages sites and let the record length of site k be m(k). Assume that the peak annual discharges for site k are given by a log normal distribution. Hence,

$$X \sim N(\mu_k, \sigma_k{}^2) \tag{5.1}$$

and the design discharge will be given exactly by

$$\text{QTRUE} = \mu_k + K(d)\sigma_k \,. \tag{5.2}$$

Now μ_k and σ_k are not known and must be estimated. Hence, the true values of μ_k and σ_k, denoted by μ_k^* and σ_k^*, must be treated as particular values of the random variables $\underline{\mu}_k$ and $\underline{\sigma}_k$. Thus, QTRUE(k) is a particular value of the random variable KQHT(k) [since by (5.2) KQHT(k) is a function of two random variables]. It is shown in Appendix A that

$$\text{QHT(k)} = \text{XBAR(k)} + t_{d,r}\left(\frac{r+2}{r+1}\right)^{\frac{1}{2}} \text{SHAT(k)} \tag{5.3}$$

gives an unbiased estimate of the design parameter since (5.3) considers the effect of sample size. In (5.3), $t_{d,r}$ is a value from tables of the Student's t at the $\frac{1}{d}$ probability level with r = degrees of freedom. The unbiased point estimate XBAR(k) and SHAT(k)^2 of μ_k and $\sigma_k^{\,2}$, respectively, were used in (4.4) to compute the normal-estimate. However, (5.3) shows that in order to compute an unbiased estimate of QTRUE(k), SHAT(k)^2 must be corrected by an amount given by

$$\left(\frac{r + 2}{r + 1}\right)^{\frac{1}{2}} t_{d,r} - K(d) .$$

By comparing tables of the Student's t with tables of the standard normal it is concluded that QHT(k) of QTRUE(k) will <u>always</u> be greater than QHN(k) for large floods. This result is demonstrated in the work of Hardison and Jennings (1972) and in the work of Benson (1952) which is discussed and explicated by Biswas (1971).

For example, let XBAR = 5.85, SHAT = .37, m(k) = 13, and d = 50. Then

$$\text{QHT(k)50} = 5.85 + (2.109)\,(\overline{14/13})\,(.6)$$

$$\text{antilog QHT(k)50} = 1513\ \text{ft}^3/\text{sec.}$$

From the previous chapter QHN(k)50 = 1210 ft/sec, and thus the normal underestimates the more exact t-estimate by $\frac{1513-1210}{1210} \cdot 100 = 25\%$.

Now let us proceed to the worth of data methodology where we must define both the cost curves and the error pdf. The hypothesized loss curves are presented in the last chapter. Let the mean of the pdf be the t-estimate QHT(k) and, as for the normal case, the pdf is approximately normal. Hence, we must define the variance of this pdf. Let the variance be denoted by Var(KQHT(k)). From (5.3)

$$\text{Var}(\text{KQHT}(k)) = \text{Var}(\underline{\mu}_k) + \left(\left(\frac{r+2}{r+1}\right)^{\frac{1}{2}} t_{d,r}\right)^2 \text{Var}(\underline{\sigma}_k) \tag{5.4}$$

Hence, Var(KQHT(k)) can be computed if $\text{Var}(\underline{\mu}_k)$ and $\text{Var}(\underline{\sigma}_k)$ can be defined. $\text{Var}(\underline{\mu}_k)$ is known to be

$$\text{Var}(\underline{\mu}_k) = \frac{\sigma_k^{\,2}}{m(k)} \tag{5.5}$$

and is statistically independent of the sample size m(k) (Hogg and Craig 1971, p. 164).

Thus, by the normality assumption, if $\text{Var}(\underline{\sigma}_k)$ could be defined, the error pdf would be defined. Since

$$\frac{\sqrt{r}\ \text{SHAT}^2}{\sigma_k^{\,2}} \sim \chi_r^2 \tag{5.6}$$

meaning that $\frac{\sqrt{r}\ \text{SHAT}^2}{\sigma_k^{\,2}}$ is distributed as Chi-square with r = degrees of freedom; then by letting $h = \left(\frac{r+2}{r+1}\right)^{\frac{1}{2}} t_{d,r}$ and rearranging (5.4)

$$\text{Var}(\text{KQHT}(k)) = \frac{\sigma_k^{\,2}}{m(k)} + \left(\frac{h\sigma_k}{r}\right)^2 \text{Var}\left(\frac{\sqrt{r}}{\sigma_k}\ \text{SHAT}(k)\right) \tag{5.7}$$

in terms of the Chi-square distributed variable. By a transformation of the Chi-square distribution, it can be shown that (see Appendix B)

$$\text{Var}(\underline{\sigma}_k) = \frac{\Psi\sigma_k^{\,2}}{r} \tag{5.8}$$

where

$$\Psi = r - 2 \left[\frac{(\frac{r+1}{2})}{(\frac{r}{2})} \right]^2 \tag{5.9}$$

and $r = m(k) - 1$.

The asymtotic value of $Var(\underline{\sigma}_k)$ in (5.8) as r becomes large is $\frac{\sigma^2}{2m(k)}$. The limiting value of Ψ as r becomes large is shown analytically to be ½ in Appendix C. Thus $Var(\underline{\sigma}_k) = \frac{\sigma_k^2}{2m(k)}$ is an approximation of $Var(\underline{\sigma}_k)$ when m(k) is large and may be called the asymptotic value of $Var(\underline{\sigma}_k)$ in (5.8). Finally, from Appendix B we find

$$Var(KQHT(k)) = \frac{\Psi t_{d,r}^2 (r+1) + r}{(r+1)r} \sigma_k^{\ 2} \tag{5.10}$$

By (5.10) it is seen that Var(QHT(k)) is a function of $\sigma_k^{\ 2}$, the population variance, which is unknown. According to Raiffa and Schlaifer (1961) the joint distribution of the unknowns $\underline{\mu}$ and $\underline{\sigma}^2$, given sample data XBAR, SHAT is the Normal-Gamma

$$f_{n\gamma}(\mu, \frac{1}{\sigma} | XBAR, SHAT, m, r)$$

and the marginal distribution of σ is called the inverted Gamma-2,

$$f_{i\gamma 2}(\sigma \mid SHAT, r) = \frac{2e^{-\frac{1}{2}r\ SHAT^2/\sigma^2} (\frac{1}{2}r\ SHAT^2/\sigma^2)^{\frac{1}{2}r+\frac{1}{2}}}{(\frac{1}{2}r-1)!\ (\frac{1}{2}r\ SHAT^2)^{\frac{1}{2}}}$$

Hence, by integrating (5.10) over σ

$$\mathrm{Var(KQHT(k))} \quad = \quad {}_0\!\int^{\infty} \frac{t^2_{d,r}(r+1) + r}{r(r+1)} \, \sigma^2 f_{iY2}(\sigma \,|\, \mathrm{SHAT}, r) d\sigma \tag{5.11}$$

it can be seen that (5.11) is proportional to the second moment of σ about the origin. Raiffa and Schlaifer (1961) give the second moment as

$$\mu_2' \quad = \quad \mathrm{SHAT}^2 \frac{r}{r-2} \quad .$$

Thus

$$\mathrm{Var(KQHT(k))} \quad = \quad \frac{\Psi t^2_{d,r}(r+1) + r}{(r-2)\ (r+1)} \, \mathrm{SHAT}^2 \ . \tag{5.12}$$

Hence, the variance of the t-estimate is a function of the sample variance SHAT^2 and the sample size $r = m(k) - 1$.

If $m(k)$ is large then in (5.12) Ψ tends to $\frac{1}{2}$, $\mathrm{SHAT}^2 \frac{r}{r-2}$ tends to σ^2, $t^2_{d,r}$ tends to $K(d)^2$, and thus

$$\mathrm{Var(KQHT(k))} \quad = \quad (\tfrac{1}{2}K(d)^2 + 1)\frac{\sigma^2}{m(k)} \quad = \quad \mathrm{Var(KQHN(k))} \tag{5.13}$$

as stated earlier (see p. 26). Hence, if $m(k)$ is large, then the effects of a small sample size have been removed.

Now we have derived the variance of the t-estimate and thus the error pdf can be shown as

$$\mathrm{KQHT(k)} \sim \mathrm{N(QHT(k)),Var(KQHT(k))} \tag{5.14}$$

To this point the cost curves have been hypothesized. The error pdf's have been defined for three types of estimates of a design parameter: (1) normal-estimate, (2) t-estimate, and (3) regression estimate. The error pdf for the regression estimate has been derived based on the variance of the residuals and the confidence interval for the conditional mean. The error pdf for two types of historical estimates has been derived.

The t-estimate and its variance have been shown to approach the normal-estimate and its variance as the sample size increases. In the next chapter an example problem is presented. The effect of the small sample size is illustrated. Also in Chapter 6, the applicability of the Bayes risk methodology to the evaluation of the worth of data is demonstrated.

CHAPTER 6

EXAMPLE OF METHODOLOGY WITH DISCUSSION

This chapter presents the results of a computerized example of the proposed methodology. We will show that two inferences may be drawn from this example. First, the historical t-estimate is significantly different from the historical normal-estimate. Furthermore, the effect of the small sample size of the data records is significant, as it creates a substantial bias in the historical normal-estimate. The second conclusion is that the Bayes risk methodology is a reliable method for pointing out the costs of error in estimation. This cost of error in estimation may be interpreted as the worth of data.

The following paragraphs describe the example and the procedure for the analysis. Then the results are discussed. The example data are taken from the plains region in Missouri. The data sites in this region were restricted to watersheds with area less than 30 square miles (as suggested by the USGS). There are 30 sites entered in the analysis. Each site is measured by six characteristics. Table 1 lists these characteristics by site. The data are supplied by the USGS (Skelton and Homyk 1970).

The analysis is performed in four iterations, the four iterations of the analysis being repeated for the 25-, 50-, and 100-year floods.

Table 1. Site index with associated topographical and climatic characteristics.

Site index	Area	Elev.	Forest cover	Precip.	2 yrs/24 hrs precipitation	Soil index
549770	2.4	.8	4.3	11.0	3.3	2.4
550200	31.0	.7	7.9	11.0	3.4	2.5
551365	3.1	.6	23.5	11.0	3.4	2.6
682000	6.0	1.1	2.0	9.0	3.3	3.2
689450	20.0	1.0	11.0	11.0	3.5	3.5
690130	.1	.9	14.8	11.0	3.3	2.4
690750	16.6	.9	7.1	13.0	3.5	2.8
691020	1.0	.8	1.0	13.0	3.4	3.5
549510	.7	.6	23.3	11.0	3.3	2.6
550300	2.6	.8	7.8	11.0	3.4	2.6
551360	1.5	.6	36.3	11.0	3.4	2.6
551420	.5	.9	1.0	11.0	3.4	2.2
681600	4.9	1.1	6.0	9.0	3.4	3.5
682030	1.3	1.1	1.0	9.0	3.3	3.2
682100	2.7	1.0	1.0	9.0	3.4	3.1
689618	.4	.9	1.0	9.0	3.3	2.4
689650	5.6	1.0	6.8	9.0	3.3	2.4
689670	.8	1.0	2.5	9.0	3.4	2.4
689720	4.7	1.0	17.4	9.0	3.3	2.4
689960	.2	.8	1.0	9.0	3.3	2.4
690250	2.5	.9	5.4	11.0	3.4	2.4
690280	1.0	.8	2.0	11.0	3.4	2.4
690470	1.0	.9	21.4	11.0	3.3	2.8
690570	.8	.7	5.6	11.0	3.4	2.8
690720	1.6	.8	1.2	13.0	3.5	2.3
690830	1.0	.8	9.0	13.0	3.5	2.8
690850	2.9	.8	2.3	13.0	3.5	2.8
690940	.3	.7	6.7	13.0	3.5	3.5
690970	.5	.8	4.0	13.0	3.5	3.5
691025	.6	.8	11.5	13.0	3.5	3.5

The First Iteration

The first iteration applies the methodology and computes the Bayes risk for the historical normal-estimate. In this iteration the normal-estimate is computed by (4.4), the variance of this estimate by (4.5), and the error pdf by (4.6). The cost curves are defined by (2.2) and (2.4) where the scaling factors a and b in (2.2) and (2.4) are chosen to be 500 and 50, respectively. These two factors are chosen such that the numbers for the Bayes risk are manageable. For the sake of argument the numbers for the Bayes risks can be called the expected utiles which are proportional to dollars per year lost because of the uncertainty in the estimate. It is not known if they actually reflect any realistic losses which might occur at the data sites. Determination of real loss functions requires another major study.

Now that the error pdf and the cost curves are defined, we can calculate the Bayes risk of the normal-estimate. The integration in (2.6) and (2.7) was performed in 20 intervals from -3.5 to +3.5 standard normal deviations from the mean of the error pdf using a Gaussian quadrature integration program (Carnahan, Luther, and Wilkes 1969). The Gaussian quadrature integration was compared with Romberg integration (Carnahan et al. 1969) and the trapezoidal rule with end point correction (Clainos 1972). The test was performed on the standard normal distribution from -3.5 to +3.5. The Gaussian quadrature program executed in half time required of the Romberg method with significance to 10^{-7}. The trapezoidal rule executed in two-thirds of the time of the Gaussian quadrature but with only three significant digits of accuracy. An increase from 20 to 100 intervals increased the accuracy by one significant digit for the Gaussian program. The Bayes risk for each of the 30 normal-estimates was then computed using the Gaussian quadrature.

The Second Iteration

The second iteration computes the Bayes risk for the t-estimate. The t-estimate is computed by (5.3), the variance of the t-estimate is computed by (5.12), and the error pdf is defined by (5.14). The cost curves which were defined for the normal-estimates are used to compute the Bayes risk for the t-estimate. Hence, the Bayes risk for the t-estimate is computed using 20 intervals and the Gaussian quadrature program.

The percentage underestimation of the normal estimate versus the t-estimate for each site can now be computed as in the example shown in Chapter 5. Figure 2 gives a histogram of the percentage difference computed by

$$\frac{QHT(k) - QHN(k)}{QHT(k)} \cdot 100 \text{ in } \%$$

for the 10-, 25-, 50-, and 100-year floods. It will be seen that the average percentage underestimation increases as the recurrence interval of the flood increases. Figure 2 illustrates the significance of the sample size in streamflow estimation. Theoretically, the t-estimate is more accurate than the normal-estimate because it compensates for the bias present in the small sample normal-estimates of return periods.

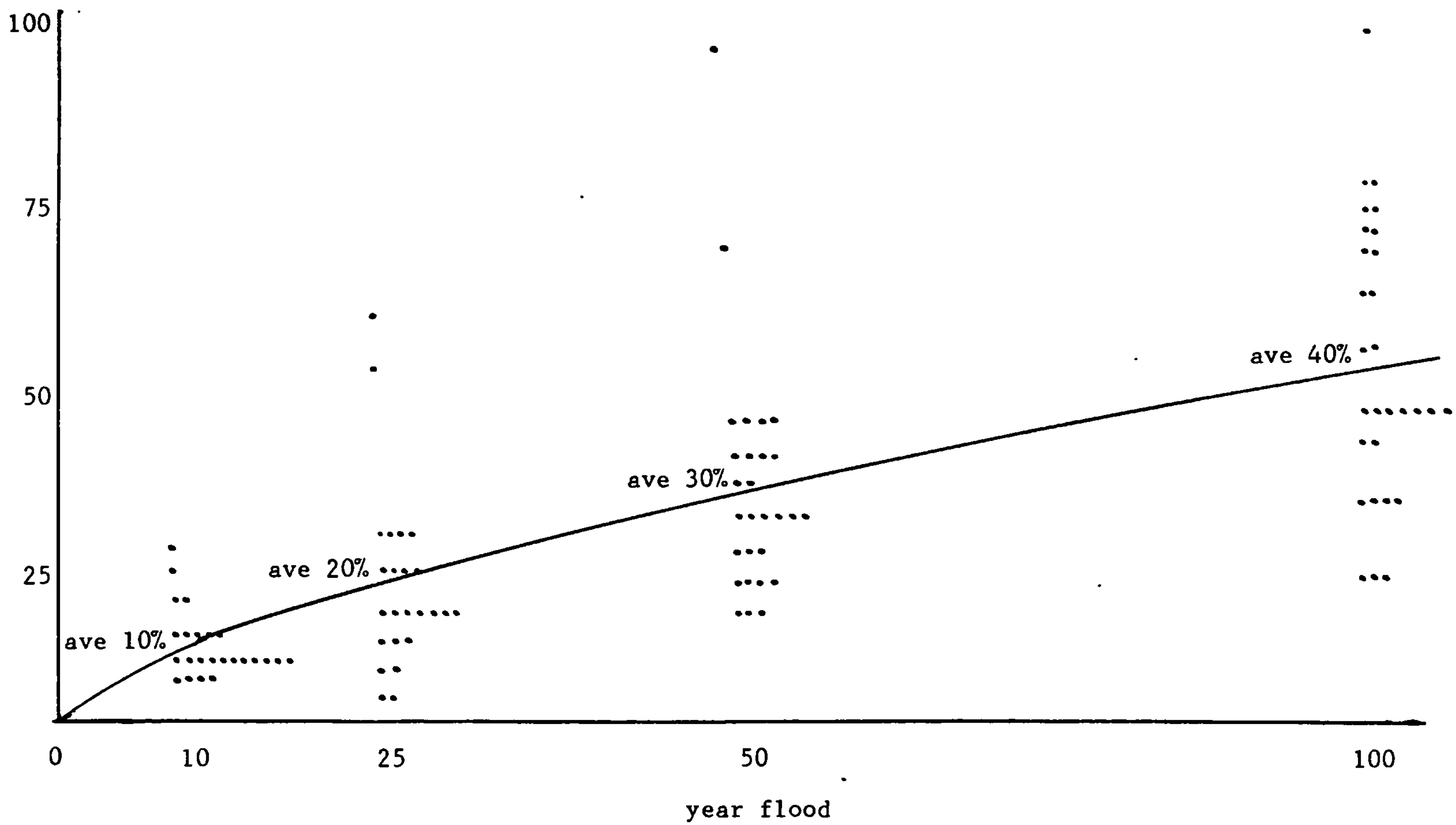

Figure 2. The increase in percentage difference between t-estimate and normal-estimate.--Each point in the bargraph represents the percentage difference between the two types of estimate for a site for a particular recurrence interval.

The difference between the two estimates is given graphically in Figure 2.

The significance of the difference between the two estimates can be stressed even further by considering the difference in the accuracy of each type of estimate. The economic consequences of the certainty in the estimates is given by the Bayes risk. The error pdf is a function of the estimate and the variance of the <u>estimate</u>. The variance of the estimate is a function of not only the sample size but also of the <u>sample</u> variance for the normal (4.5) and for the t (5.12).

The significance of the small sample bias in the normal-estimate is demonstrated in the following analysis. First, the Bayes risk of the historical t-estimate is computed (Figure 3). Next, the Bayes risk of the historical normal-estimate is computed <u>based on</u> the error pdf of the t-estimate, which is an unbiased estimate (Figure 4). The difference between the two Bayes risks is the extra expected loss one suffers in ignoring the influence of the small sample bias. These values are summarized in Tables 2, 3, and 4 for the 25-, 50-, and 100-year floods. Notice in Table 2 that for the 25-year flood there is a difference of 33% based on the Bayes risk of the normal-estimate using the error pdf for the t-estimate. This means that if the decision maker uses the normal-estimate as the design parameter, then he is suffering on the average a penalty of 33% in Bayes risk for not considering the impact of the small sample size. For the 50-year flood this penalty increases to about 40% and for the 100-year flood the penalty is about 44%.

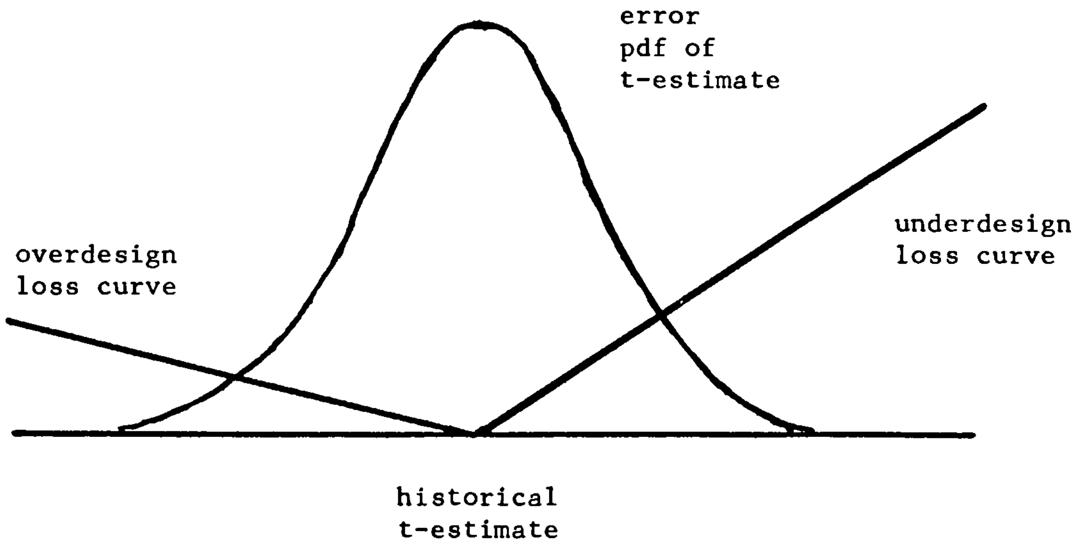

Figure 3. The Bayes risk of the t-estimate.

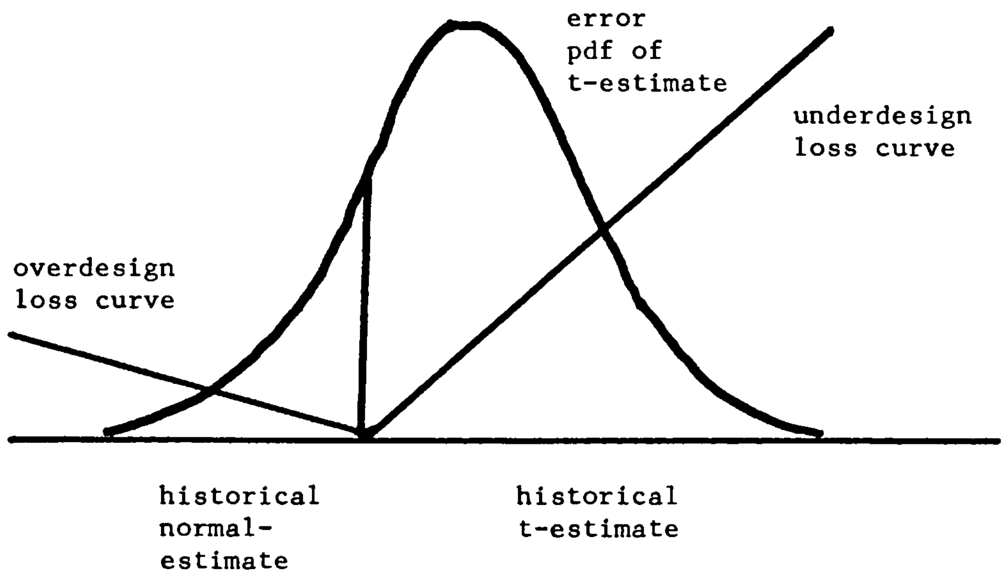

Figure 4. The Bayes Risk of the normal-estimate based on the error pdf of the t-estimate.

Table 2. Bayes risk of t-estimate vs. normal-estimate for 25-year flood.

Site	Coefficient of variation	Log drainage area	Risk in using t-estimate	Risk in using normal estimate (t-error pdf)	Diff and percent	
690130	.073	-2.04	57.	84.	28.	33.
689960	.466	-1.56	280.	417.	138.	33.
690940	.126	-1.20	96.	149.	53.	36.
689618	.223	-.97	157.	235.	77.	33.
690970	.329	-.71	232.	346.	114.	33.
551420	.118	-.62	103.	158.	55.	35.
691025	.116	-.60	116.	180.	64.	36.
549510	.201	-.36	141.	213.	72.	34.
690570	.139	-.22	107.	159.	53.	33.
689670	.151	-.22	126.	188.	62.	33.
690470	.187	-.04	156.	233.	77.	33.
690830	.108	-.03	92.	137.	45.	33.
691020	.137	.01	111.	165.	54.	33.
690280	.208	.04	159.	237.	78.	33.
682030	.100	.26	80.	117.	37.	31.
551360	.181	.41	144.	214.	71.	33.
690720	.113	.50	102.	152.	50.	33.
549770	.105	.87	88.	131.	43.	33.
690250	.079	.92	75.	114.	38.	34.
550300	.059	.97	58.	87.	29.	34.
682100	.179	1.00	132.	187.	55.	30.
690850	.083	1.05	76.	113.	37.	33.
551365	.153	1.12	137.	205.	67.	33.
689720	.136	1.55	139.	208.	68.	33.
681600	.158	1.59	118.	167.	49.	30.
689650	.151	1.72	148.	223.	76.	34.
682000	.167	1.80	124.	174.	50.	29.
690750	.090	2.81	93.	137.	44.	32.
689450	.130	3.00	120.	171.	52.	30.
550200	.105	3.43	76.	102.	27.	26.

AVERAGE PERCENT DIFFERENCE 33.0
STND ERROR PERCENT DIFFERENCE 2.0

Table 3. Bayes risk of t-estimate vs. normal-estimate for 50-year flood.

Site	Coefficient of variation	Log drainage area	Risk in using t-estimate	Risk in using normal estimate (t-error pdf)	Diff and Percent	
690130	.073	-2.04	65.	106.	41.	39.
689960	.466	-1.56	319.	523.	204.	39.
690940	.126	-1.20	111.	191.	80.	42.
689618	.223	-.97	180.	294.	115.	39.
690970	.329	-.71	265.	434.	169.	39.
551420	.118	-.62	119.	200.	82.	41.
691025	.116	-.60	134.	230.	96.	42.
549510	.201	-.36	162.	269.	107.	40.
690570	.139	-.22	122.	200.	78.	39.
689670	.151	-.22	144.	235.	92.	39.
690470	.187	-.04	178.	292.	114.	39.
690830	.108	-.03	105.	172.	67.	39.
691020	.137	.01	126.	207.	80.	39.
690280	.208	.04	181.	297.	116.	39.
682030	.100	.26	91.	145.	54.	37.
551360	.181	.41	164.	269.	105.	39.
690720	.113	.50	116.	190.	74.	39.
549770	.105	.87	101.	165.	64.	39.
690250	.079	.92	86.	144.	57.	40.
550300	.059	.97	66.	110.	44.	40.
682100	.179	1.00	149.	229.	81.	35.
690850	.083	1.05	87.	142.	55.	39.
551365	.153	1.12	157.	257.	100.	39.
689720	.136	1.55	159.	260.	101.	39.
681600	.158	1.59	133.	206.	72.	35.
689650	.151	1.72	169.	282.	112.	40.
682000	.167	1.80	140.	213.	73.	34.
690750	.090	2.81	106.	171.	65.	38.
689450	.130	3.00	136.	211.	76.	36.
550200	.105	3.43	85.	124.	39.	31.

AVERAGE PERCENT DIFFERENCE 38.0
STND ERROR PERCENT DIFFERENCE 2.0

Table 4. Bayes risk of t-estimate vs. normal-estimate for 100-year flood.

Site	Coefficient of variation	Log drainage area	Risk in using t-estimate	Risk in using normal estimate (t-error pdf)	Diff and Percent	
690130	.073	-2.04	73.	131.	58.	44.
689960	.466	-1.56	359.	645.	286.	44.
690940	.126	-1.20	126.	239.	113.	47.
689618	.223	-.97	202.	363.	161.	44.
690970	.329	-.71	299.	536.	238.	44.
551420	.118	-.62	134.	250.	116.	46.
691025	.116	-.60	152.	288.	136.	47.
549510	.201	-.36	182.	333.	151.	45.
690570	.139	-.22	137.	247.	109.	44.
689670	.151	-.22	162.	290.	129.	44.
690470	.187	-.04	201.	360.	160.	44.
690830	.108	-.03	118.	212.	94.	44.
691020	.137	.01	142.	255.	113.	44.
690280	.208	.04	204.	367.	163.	44.
682030	.100	.26	102.	178.	76.	43.
551360	.181	.41	185.	332.	147.	44.
690720	.113	.50	131.	235.	104.	44.
549770	.105	.87	113.	203.	90.	44.
690250	.079	.92	97.	178.	80.	45.
550300	.059	.97	75.	136.	62.	45.
682100	.179	1.00	166.	278.	112.	40.
690850	.083	1.05	97.	175.	78.	44.
551365	.153	1.12	176.	317.	140.	44.
689720	.136	1.55	179.	321.	142.	44.
681600	.158	1.59	149.	249.	100.	40.
689650	.151	1.72	191.	349.	158.	45.
682000	.167	1.80	156.	258.	102.	40.
690750	.090	2.81	119.	210.	91.	43.
689450	.130	3.00	152.	257.	105.	41.
550200	.105	3.43	94.	148.	53.	36.

AVERAGE PERCENT DIFFERENCE 44.0
STND ERROR PERCENT DIFFERENCE 2.0

In other words, although the design specification was chosen to be optimal, the decision maker, by using the normal-estimate, is not designing the bridge or culvert according to his optimal specification: his actual design is indeed <u>not</u> optimal. The decision maker suffers a risk in designing upon any estimate of the optimal design specification, but the risk he suffers by using the normal-estimate is <u>not</u> the risk arising by use of the estimate of the optimal design parameter. Only by using the t-estimate which considers the small sample size is the decision maker implementing his optimal decision.

The Third Iteration

The third iteration in the analysis of the example involved regression and was accomplished in two parts. This was a forward stepwise regression of the six variables in Table 1 on the design flood with an entry F level of 0.0. Thus, all variables were included. This regression analysis is not in strict compliance with USGS procedures but only serves to illustrate the methodology.

The first part of the third iteration was a regression based on the normal-estimates as inputs. The second part was a regression based on the t-estimates as inputs. In each of these regressions one site was excluded and the regression involved the data sets of the other 29 sites. This exclusion of a site was repeated for each of the 30 sites. The regression equation computed on the basis of the set of 29 sites was then used to compute a <u>prediction</u> of the log of the flow for the thirtieth site. Prediction rather than estimation is used here to indicate that the site was omitted from the regression. Thus, for the normal-inputs there are 30 regression equations and for the t-inputs there are 30 regression equations.

For the normal case, then, the error pdf for the prediction of the log of the flow at the excluded site is given by (3.10) with mean (3.12) and variance (3.22). The cost curves for the sites are as described earlier in this chapter. Thus the Bayes risk of the regression prediction based on the normal inputs can be computed for each of the 30 sites, each site being excluded in separate regressions. This procedure was repeated for the case of the t-inputs and tabulated in Tables 5, 6, and 7.

The Bayes risks of the regression prediction based on the normal inputs can be contrasted with the regression prediction based on the t-inputs in the same manner as presented for the historical case. The conclusion from this contrast is the same as from the contrast of the uncertainty of the normal-estimate and the t-estimate. The uncertainty is measured by Bayes risk of the regression prediction based on the t-estimate.

It should be stressed here that the so-called "regression line" is reoriented when based on the t-inputs. For the 25-year flood the normal-estimate is 421 cfs while the regression prediction based on the normal-inputs excluding site 690130 is 374 cfs, a difference of -12%. For the 25-year flood the t-estimate is 462 ft^3sec while the regression prediction based on the t-inputs excluding site 690130 is 495 ft^3/sec, a difference of +6.5%. In other words the normal regression line overestimates the historical normal-estimate while the

Table 5. Comparison of historical and regression normal- and t-estimates for 25-year flood.

Normal-estimate		t-Estimate		All sites included
Hist	Regr	Hist	Regr	Regr t-estimate
421.	374.	462.	495.	481.
1951.	428.	3075.	537.	875.
200.	834.	238.	1088.	746.
917.	969.	1184.	1254.	1240.
2266.	640.	3307.	761.	1004.
664.	967.	795.	1179.	1084.
935.	868.	1152.	1063.	1082.
505.	906.	641.	1150.	970.
776.	1099.	924.	1407.	1347.
1495.	2001.	1834.	2550.	2313.
2192.	930.	2825.	1067.	1405.
1153.	1082.	1339.	1291.	1298.
1065.	919.	1274.	1115.	1178.
1409.	1206.	1825.	1469.	1499.
1376.	2669.	1556.	3428.	2843.
1446.	1328.	1827.	1679.	1718.
1841.	1120.	2172.	1303.	1503.
1016.	1828.	1173.	2167.	1910.
1241.	2251.	1409.	2645.	2494.
1243.	2128.	1369.	2540.	2350.
4277.	3543.	5150.	4531.	4668.
1407.	1960.	1592.	2303.	2182.
2678.	1753.	3347.	2184.	2436.
6729.	3523.	8440.	4090.	4796.
3838.	6756.	4534.	8394.	7098.
3356.	4554.	4304.	5304.	5134.
5364.	5114.	6360.	6151.	6203.
6278.	4799.	7271.	5145.	5631.
12543.	8077.	14920.	9121.	10510.
8487.	6255.	9304.	7421.	8011.

Table 6. Comparison of historical and regression normal- and t-estimates for 50-year flood.

Normal-estimate Hist	Normal-estimate Regr	t-Estimate Hist	t-Estimate Regr	All sites included Regr t-estimate
474.	476.	541.	714.	638.
3523.	572.	6752.	791.	1440.
236.	1157.	303.	1694.	1105.
1278.	1363.	1843.	1970.	1945.
3703.	785.	6356.	1008.	1426.
805.	1227.	1043.	1631.	1483.
1142.	1116.	1545.	1492.	1503.
669.	1219.	941.	1717.	1441.
973.	1513.	1248.	2154.	2036.
1950.	2757.	2612.	3899.	3463.
3048.	1087.	4382.	1323.	1856.
1400.	1313.	1735.	1694.	1700.
1345.	1214.	1739.	1602.	1657.
1972.	1552.	2854.	2060.	2122.
1660.	3932.	1976.	5615.	4384.
1959.	1784.	2737.	2497.	2560.
2283.	1330.	2892.	1652.	1933.
1224.	2308.	1503.	2943.	2564.
1442.	2786.	1730.	3509.	3285.
1394.	2668.	1603.	3438.	3124.
6051.	5166.	7872.	7335.	7456.
1652.	2438.	1971.	3070.	2878.
3579.	2374.	4925.	3250.	3614.
9031.	4449.	12487.	5504.	6590.
5237.	9552.	6630.	13010.	10829.
4507.	5880.	6440.	7303.	7161.
7518.	6979.	9567.	9072.	9193.
7725.	5455.	9527.	6021.	6788.
17005.	10142.	21758.	12050.	14285.
10838.	8120.	12331.	10359.	10988.

Table 7. Comparison of historical and regression normal- and t-estimates for 100-year flood.

Normal-estimate Hist	Normal-estimate Regr	t-Estimate Hist	t-Estimate Regr	All sites included Regr t-estimate
528.	592.	631.	1029.	845.
5978.	740.	14474.	1151.	2335.
274.	1550.	385.	2603.	1621.
1721.	1850.	2830.	3050.	3006.
5744.	943.	11972.	1325.	2006.
957.	1519.	1361.	2237.	2013.
1366.	1398.	2065.	2073.	2071.
861.	1590.	1369.	2534.	2118.
1191.	2013.	1670.	3255.	3039.
2473.	3672.	3680.	5879.	5119.
4094.	1249.	6705.	1633.	2434.
1666.	1562.	2229.	2210.	2213.
1658.	1558.	2351.	2270.	2303.
2663.	1945.	4402.	2858.	2974.
1962.	5561.	2485.	9011.	6641.
2571.	2322.	4049.	3670.	3769.
2766.	1552.	3816.	2085.	2469.
1446.	2844.	1910.	3956.	3407.
1650.	3372.	2113.	4612.	4289.
1545.	3267.	1868.	4612.	4118.
8253.	7239.	11766.	11640.	11669.
1907.	2964.	2423.	4054.	3762.
4639.	3114.	7160.	4770.	5292.
11752.	5483.	18254.	7311.	8940.
6917.	13021.	9503.	19780.	16201.
5867.	7391.	9532.	9907.	9848.
10171.	9217.	14069.	13138.	13365.
9301.	6118.	12356.	6990.	8112.
22328.	12433.	31128.	15688.	19110.
13490.	10255.	16045.	14257.	14838.

t-regression line underestimates the historical t-estimate for this example. The t-estimate for the large design floods is always greater than the normal-estimate and, hence, the t-regression line will always have greater y-intercept and slightly different slope than the normal regression line. Specifically, if site 690130 is excluded from the regression for the 25-year flood, then not only is the t-regression line reoriented higher (495 versus 374) but is so high that it overestimates the historical estimate. Theoretically, the t-estimate is more accurate than the normal, and hence the t-regression line should define more accurate predictions.

Here again the significance of the small sample bias is strikingly demonstrated. First, the Bayes risk of the regression estimate (when the regression is performed on the t-inputs and excludes site k) is computed (see Figure 5). Next, the Bayes risk of the regression estimate (when the regression is performed on the normal inputs and excludes site k) is computed based on the error pdf of the regression t-estimate (see Figure 6). The difference in the two Bayes risks is the extra expected loss one suffers in ignorance of the influence of the small sample bias when carried through the regression analysis. These values are summarized in Tables 8, 9, and 10 for the 25-, 50-, and 100-year floods. These tables show that on the average the decision maker is suffering a penalty of about 24% for the 25-year flood, 28% for the 50-year flood, and 31% for the 100-year flood. These results indicate that the influence of the sample size upon the regression predictions, although smaller than for the historical estimates, is still substantial.

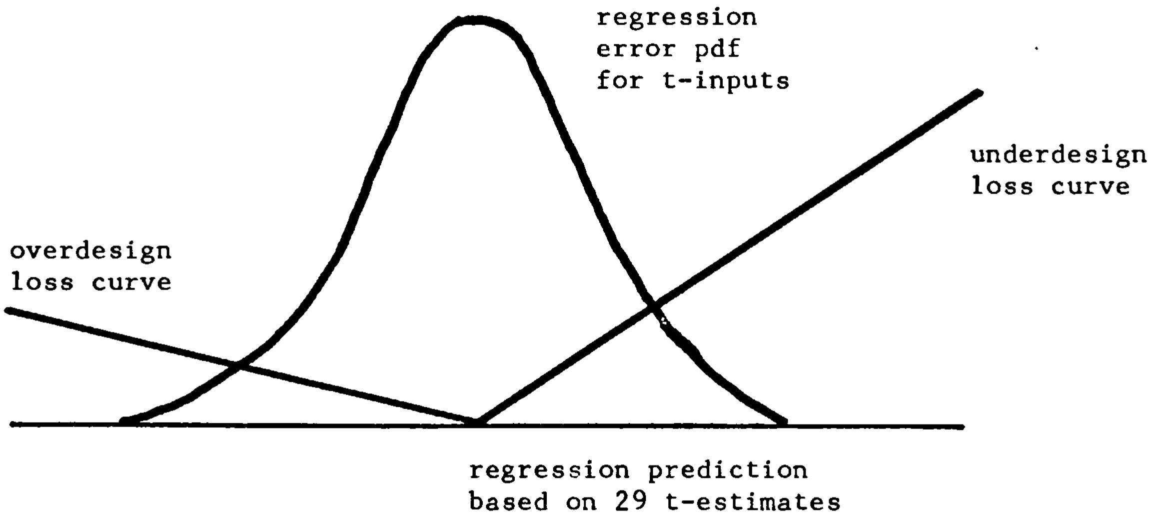

Figure 5. Bayes risk of regression prediction based on t-estimates.

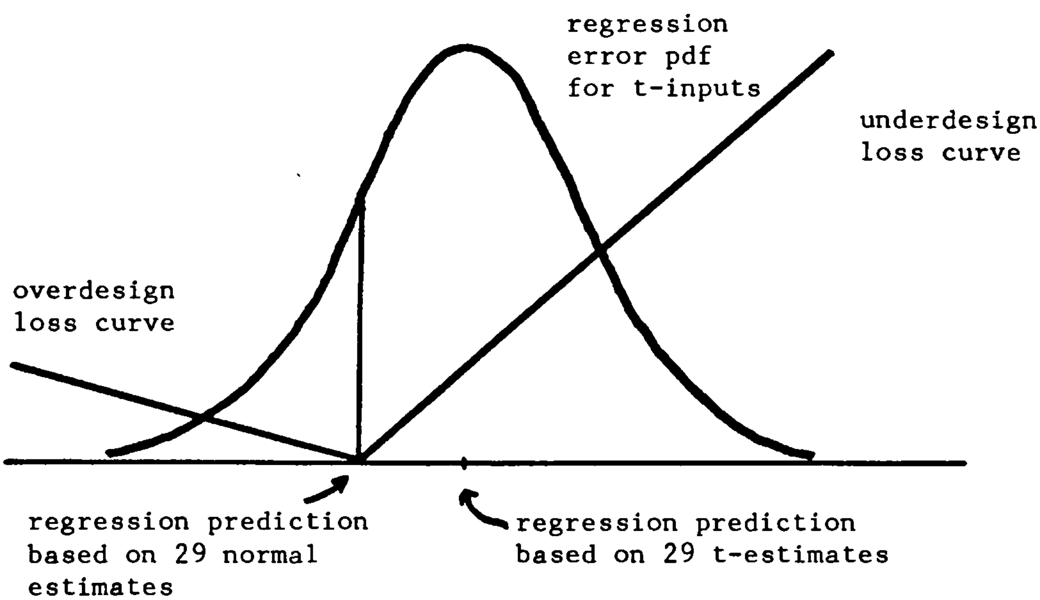

Figure 6. Bayes risk of regression prediction using normal-estimates as inputs with regression error pdf of t-inputs.

Table 8. Bayes risk for regression estimate for 25-year flood.

Regr risk using t-estimate	Regr risk using regr normal-estimate with regr t-error pdf	Difference	Percent
206.	289.	83.	29.
158.	225.	67.	30.
161.	242.	81.	33.
177.	253.	77.	30.
155.	205.	50.	24.
178.	235.	57.	24.
180.	238.	58.	24.
186.	256.	70.	27.
166.	240.	74.	31.
188.	259.	71.	27.
179.	217.	38.	18.
172.	223.	51.	23.
207.	262.	55.	21.
166.	224.	57.	26.
176.	250.	74.	30.
186.	255.	68.	27.
185.	227.	42.	19.
175.	223.	48.	22.
163.	209.	46.	22.
166.	217.	51.	23.
181.	254.	73.	29.
171.	216.	46.	21.
183.	246.	64.	26.
175.	217.	42.	19.
183.	246.	63.	26.
173.	216.	43.	20.
184.	236.	53.	22.
184.	202.	19.	9.
186.	220.	33.	15.
195.	243.	48.	20.

AVERAGE PERCENT DIFFERENCE 24.0
STND ERROR PERCENT DIFFERENCE 5.0

Table 9. Bayes risk for regression estimate for 50-year flood.

Regr risk using t-estimate	Regr risk using regr normal-estimate with regr t-error pdf	Difference	Percent
246.	369.	123.	33.
187.	286.	99.	35.
196.	314.	118.	38.
211.	324.	112.	35.
182.	256.	74.	29.
213.	296.	84.	28.
215.	301.	85.	28.
223.	326.	103.	32.
198.	306.	108.	35.
225.	329.	104.	32.
214.	269.	56.	21.
206.	280.	74.	26.
247.	327.	80.	24.
199.	283.	84.	30.
209.	318.	108.	34.
223.	323.	101.	31.
222.	283.	62.	22.
210.	280.	70.	25.
195.	262.	67.	25.
198.	272.	74.	27.
217.	323.	106.	33.
204.	270.	66.	25.
219.	312.	93.	30.
210.	271.	61.	22.
219.	311.	92.	29.
207.	269.	62.	23.
219.	295.	76.	26.
220.	246.	27.	11.
223.	271.	48.	18.
233.	303.	70.	23.

AVERAGE PERCENT DIFFERENCE 28.0
STND ERROR PERCENT DIFFERENCE 6.0

Table 10. Bayes risk for regression estimate for 100-year flood.

Regr risk using t-estimate	Regr risk using regr normal-estimate with regr t-error pdf	Difference	Percent
285.	457.	172.	38.
216.	354.	138.	39.
230.	395.	165.	42.
246.	402.	156.	39.
210.	313.	103.	33.
247.	364.	116.	32.
250.	369.	119.	32.
260.	403.	143.	35.
230.	381.	151.	40.
262.	406.	144.	36.
248.	325.	77.	24.
239.	343.	103.	30.
288.	398.	110.	28.
231.	347.	117.	34.
242.	392.	151.	38.
259.	399.	140.	35.
258.	343.	85.	25.
244.	342.	97.	28.
228.	320.	93.	29.
231.	334.	103.	31.
252.	399.	147.	37.
237.	329.	92.	28.
255.	384.	129.	34.
245.	328.	83.	25.
256.	382.	126.	33.
241.	326.	85.	26.
255.	360.	105.	29.
255.	291.	36.	12.
259.	324.	66.	20.
272.	367.	96.	26.

AVERAGE PERCENT DIFFERENCE 31.0
STND ERROR PERCENT DIFFERENCE 6.0

If the decision maker uses the regression estimate of the design parameter when the regression is based on the t-inputs, then he will be implementing his optimal design decision.

The Fourth Iteration

The fourth iteration performed in the example analysis also involves regression. In this iteration all 30 sites were included in the regression analysis and the inputs were the t-estimates. For each of the 30 sites the Bayes risk for the regression estimate was computed. The error pdf's for this iteration are defined by (3.21) with mean (3.10) and variance (3.22). The cost curves remain as defined earlier. The difference between this iteration and the third iteration is that this time all sites are included and, thus, only one regression equation results from the analysis for each of the three design floods. Hence, the Bayes risk for site k computed in the fourth iteration is the cost of the uncertainty in using the regression estimate for site k when site k has helped to determine the regression coefficients. The emphasis here is not to determine an estimate of the design parameter but to evaluate the contribution of each data set to the regression equation.

Thus, the difference between the Bayes risk for site k when site k has been excluded from the regression and the Bayes risk for site k when all sites are included in the regression (now to be called the differential Bayes risk) gives some measure of the value of site k in the regionalization process. That is, how much is the Bayes risk of the regression prediction reduced when the site data set is included in the regression? Tables 11, 12, and 13 give the differential Bayes risk for the 25-, 50-, and 100-year floods. It will be observed that site 690130 stands out as having the greatest differential Bayes risk, relative to the differential Bayes risk values for the other sites. That is, a regression prediction for this site will incur the greatest positive differential Bayes risk or the most uncertainty in the estimate. Since the Bayes risk is formulated on the basis of confidence intervals, one might hypothesize this conclusion since this site is at the extreme lower limit (where the confidence limits are widest) of the interval of values over which we might try to predict by regression. The most significant independent variable in terms of explaining the variance of the dependent variable is the area, and 690130 has the smallest area. However, note that site 691020 has the next greatest positive value of the differential Bayes risk. The area for this site lies approximately in the center of the interval of site areas. Now observe the Bayes risk in Table 11 for the site with the largest area (the other extreme), site 550200. This value of Bayes risk is not outstanding. Thus, the Bayes risk procedure dramatically demonstrates those site data sets which potentially will contribute most to defining more accurate predictions by regression. That is, more data sets like the data set for site 690130 and 691020 should be included in the regression.

Now since a positive differential Bayes risk in Tables 11, 12, and 13 gives the amount of Bayes risk which is reduced by the inclusion of a data set in the regression, then a negative differential

Table 11. Potential worth of data for 25-year flood.

Site	Coefficient of variation	Log area	Risk in using regr estimate (t-estimate, all sites included)	Risk in using regr prediction (t-estimate, site excluded)	Diff
690130	.073	-2.04	184.	206.	22.
689960	.466	-1.56	176.	158.	-18.
690940	.126	-1.20	174.	161.	-13.
689618	.223	-.97	170.	177.	7.
690970	.329	-.71	169.	155.	-15.
551420	.118	-.62	171.	178.	7.
691025	.116	-.60	172.	180.	8.
549510	.201	-.36	177.	186.	9.
690570	.139	-.22	163.	166.	3.
689670	.151	-.22	177.	188.	11.
690470	.187	-.04	176.	179.	3.
690830	.108	-.03	167.	172.	6.
691020	.137	.01	185.	207.	22.
690280	.208	.04	162.	166.	4.
682030	.100	.26	173.	176.	3.
551360	.181	.41	175.	186.	11.
690720	.113	.50	176.	185.	9.
549770	.105	.87	171.	175.	4.
690250	.079	.92	162.	163.	0.
550300	.059	.97	165.	166.	1.
682100	.179	1.00	173.	181.	9.
690850	.083	1.05	166.	171.	4.
551365	.153	1.12	174.	183.	8.
689720	.136	1.55	172.	175.	3.
681600	.158	1.59	175.	183.	8.
689650	.151	1.72	167.	173.	5.
682000	.167	1.80	174.	184.	10.
690750	.090	2.81	175.	184.	9.
689450	.130	3.00	176.	186.	10.
550200	.105	3.43	180.	195.	15.

Table 12. Potential worth of data for 50-year flood.

Site	Coefficient of variation	Log area	Risk in using regr estimate (t-estimate, all sites included)	Risk in using regr prediction (t-estimate, site excluded)	Diff
690130	.073	-2.04	220.	246.	26.
689960	.466	-1.56	210.	187.	-23.
690940	.126	-1.20	208.	196.	-12.
689618	.223	-.97	203.	211.	8.
690970	.329	-.71	203.	182.	-20.
551420	.118	-.62	205.	213.	8.
691025	.116	-.60	205.	215.	10.
549510	.201	-.36	211.	223.	12.
690570	.139	-.22	195.	198.	3.
689670	.151	-.22	212.	225.	14.
690470	.187	-.04	210.	214.	3.
690830	.108	-.03	199.	206.	7.
691020	.137	.01	221.	247.	27.
690280	.208	.04	194.	199.	4.
682030	.100	.26	207.	209.	2.
551360	.181	.41	210.	223.	13.
690720	.113	.50	210.	222.	11.
549770	.105	.87	204.	210.	6.
690250	.079	.92	194.	195.	1.
550300	.059	.97	197.	198.	1.
682100	.179	1.00	206.	217.	11.
690850	.083	1.05	199.	204.	5.
551365	.153	1.12	208.	219.	11.
689720	.136	1.55	205.	210.	5.
681600	.158	1.59	210.	219.	10.
689650	.151	1.72	200.	207.	7.
682000	.167	1.80	208.	219.	12.
690750	.090	2.81	209.	220.	11.
689450	.130	3.00	211.	223.	12.
550200	.105	3.43	215.	233.	18.

Table 13. Potential worth of data for 100-year flood.

Site	Coefficient of variation	Log area	Risk in using regr estimate (t-estimate, all sites included)	Risk in using regr prediction (t-estimate, site excluded)	Diff
690130	.073	-2.04	256.	285.	29.
689960	.466	-1.56	244.	216.	-29.
690940	.126	-1.20	241.	230.	-11.
689618	.223	-.97	236.	246.	10.
690970	.329	-.71	236.	210.	-25.
551420	.118	-.62	238.	247.	9.
691025	.116	-.60	239.	250.	12.
549510	.201	-.36	246.	260.	15.
690570	.139	-.22	227.	230.	3.
689670	.151	-.22	246.	262.	16.
690470	.187	-.04	245.	248.	3.
690830	.108	-.03	232.	239.	8.
691020	.137	.01	257.	288.	31.
690280	.208	.04	226.	231.	5.
682030	.100	.26	240.	242.	2.
551360	.181	.41	244.	259.	15.
690720	.113	.50	244.	258.	14.
549770	.105	.87	237.	244.	7.
690250	.079	.92	226.	228.	2.
550300	.059	.97	229.	231.	1.
682100	.179	1.00	240.	252.	12.
690850	.083	1.05	231.	237.	6.
551365	.153	1.12	242.	255.	13.
689720	.136	1.55	239.	245.	6.
681600	.158	1.59	244.	256.	12.
689650	.151	1.72	232.	241.	8.
682000	.167	1.80	242.	255.	14.
690750	.090	2.81	243.	255.	12.
689450	.130	3.00	245.	259.	14.
550200	.105	3.43	250.	272.	22.

Bayes risk in Tables 11, 12, and 13 gives the amount of Bayes risk which was added by the inclusion of a data set. That is, the uncertainty in a regression prediction has been increased by the inclusion of a data set. Notice that the second site 689960 in Tables 11, 12, and 13 has a fairly high negative differential Bayes risk. This negative value represents the increased uncertainty in the regression equation to predict the flow for a site with the characteristics of 689960. However, the risk for 689960 does not mean that 689960 should be deleted from the regression. The unusually high sample variance (3.1) for this site probably accounts for its disturbing effect upon the regression equation when it is included. The record for 689960 should be continued until the sample variance "settles down." Obviously according to the Bayes risks in Tables 11, 12, and 13, more data sets similar to that of 690130, 691020, and 689960 should be collected and included in the regression analysis. The potential worth of such data sets is given by the Bayes risk for the indicated sites since this is the differential risk caused by the inclusion of a particular set of data values. There is a possibility also that site 690130 should not be included in the regression equation for this region. That is, geographically it may lie on the border separating two regions. The differential Bayes risk approach can then show to which region such a site should belong by examining the risk of the prediction for the site as computed by the two regional regression equations.

In summary, the t-estimate provides a significantly different estimate of a design parameter than the normal. Furthermore, there can be a substantially greater risk associated with the use of the t-estimate. The regression predictions are changed when the t-estimates are used as input. The Bayes risk of an estimate is a measure of the worth of data and can be used to indicate dramatically where collection emphasis might be made.

CHAPTER 7

DISCUSSION AND CONCLUSIONS

It has already been pointed out that the estimate of the optimal design parameter should be computed from the Student's t-distribution. For the floods of higher recurrence interval (25, 50, and 100 years) the percentage difference between the estimate computed from the normal (an approximation) and that computed from the t-distribution is significant. It is important to use the t-estimates as inputs to the regression analysis so that, on the average, the regression predictions will be unbiased in the sense that the small sample bias is not carried by the historical inputs to the regression prediction outputs.

If a log normal distribution of floods is assumed, then all available records may be used in computing the regression equation by using the t-estimate. For example, Thomas and Benson (1970) remark that in order to include an estimate of the 50-year flood into the regression, the estimate should be computed from a record length of 25 years (half the recurrence interval). The t-estimate, because it considers sample size, would permit all records to be included into the regression.

The Bayes risk of the historical estimate is computed by integrating out the uncertainty in the population parameters given that the decision about the design parameter is indeed optimal.

The differential risk in excluding versus including a historical estimate in the regression model is a measure of the worth of data given that the region is homogeneous and that the regression model is appropriate. The differential risk of a site when contrasted with the differential risk of other sites reflects the relative amount of perturbation to the regional regression equation. A high positive differential risk indicates that the inclusion of a site into the regression has relatively improved the predictive power of the regression equation. Hence, more data analogous to that particular data set are suggested.

A high negative differential risk indicates that the exclusion of a site from the regression has relatively improved the predictive power of the regression equation. Hence, more data analogous to that particular data set are suggested. Consequently if differential Bayes risk is high in absolute value for a site, analogous sites should be included into the regression. Thus, the differential risk approach can be used to quantitatively assess the relative worth of data for the regionalization process.

Suggestions for Further Research

In the Introduction, we were interested in defining a methodology for evaluating the payoff of collecting more data. Management would then trade off the cost of collecting the data against the payoff of this data. The differential Bayes risk is suggested in this report to be a pre'iminary approach to quantifying the payoff of more data. It has been shown that the differential Bayes risk as described in the preceding chapter can be interpreted as the worth of data for regionalization.

A second theme of this report is to stress the importance of considering the small sample size as it affects the estimation of a decision design parameter. The unifying thread connecting the worth of data and the effects of small sample size is the economic consequences of the decisions made concerning uncertain outcomes in nature.

Several suggestions for further research clearly arise out of this investigation. The research with the highest priority perhaps is the definition of some realistic loss curves for errors in estimation. A study should be made of the physical economic consequences of faulty design given an optimal decision. An outcome of such a study should be generalized cost curves for a region. Such generalized cost curves would permit the computation of an expected differential Bayes risk. This expected differential Bayes risk could then be analyzed according to specific variables and ranges of the variables which potentially contribute the most "information" to the regression equation. Exclusion of more than one site should be considered but this immediately leads into combinational problems. It may be suggested that the residual variance might serve as the variance of the error pdf for the regression prediction instead of a variance based on a confidence interval principle. What are the effects of this substitution? Parenthetically, can the regression equation be computed by minimizing the risk instead of minimizing the residual variance? The Bayes risk approach can quantify the worth of data and more work on this approach should be pursued once realistic cost curves for a region can be generated.

Of secondary importance to the generation of cost curves is an analysis of the economic consequences of estimating from a small sample size. Since the risk is a function of the size of the estimate and since for large design floods the t-estimate is <u>always</u> larger than the normal-estimate, then, in a worth of data study, how important is the sample size? This question is especially relevant in working with a skewed distribution like log Pearson III. When evaluating the worth of data by computing the Bayes risk, it is important to use an estimate which accounts for the effects of bias in a small sample size.

Another question to be answered here is the relative size of the risk in estimation with the risk of the optimal design. It has been shown in this report that the sample size can create a substantial bias in the estimate which is computed by a technique which does not account for the uncertainty in the population parameters.

Conclusions

The Bayes risk approach to the evaluation of the worth of data is considered to be an extension and improvement over the variance approach. The Bayes risk approach considers the costs of error in estimation. These costs of error may be different for each site due to the size of the structure. Hence, the Bayes risk approach customizes the costs of error as given by the variance of the estimate for each site. Since the t-estimate reflects the bias of a small sample size, the costs of error in estimation may be quite different from the normal-estimate costs.

1. The first conclusion of this thesis is that the t-distribution should be used to compute the log of the flood when a log-normal distribution has been assumed because the bias of a small sample size is then removed. Furthermore, the use of the t-estimate allows all data on hand to enter the regionalization equation: the cost of the uncertainty in the regression prediction can be measured by the Bayes risk of the prediction.

2. The second conclusion of this thesis is that the differential Bayes risk is a measure of the worth of data for regionalization. By defining a set of generalized loss functions for a region, the data collection effort can eventually be optimized by minimizing the differential Bayes risk in design parameter prediction.

APPENDIX A

THE t-ESTIMATE

It can be shown (Hogg and Craig 1971) that

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

where

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

and $x_i = \log y_i$ for y_i the ith year maximum annual flow. Now since

$$\hat{s}^2 = \frac{\Sigma(x_i - \bar{x})}{n - 1}$$

then

$$\frac{(n-1)\hat{s}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Letting r = n-1 (the degrees of freedom), then

$$\frac{r\hat{s}^2}{\sigma^2} \sim \chi^2_{(r)}$$

Since

$$X \sim N(\mu, \sigma^2)$$

and

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

then

$$\bar{U} = X - \bar{X} \sim N(0, \sigma^2 + \frac{\sigma^2}{n}) \qquad (A.1)$$

and the standard normalizing $\bar{U}$ to Z

$$Z \sim N(0,1) .$$

Thus from (A.1) $Z = z_i = \dfrac{x_i - \bar{x} - 0}{\sqrt{\sigma^2 + \dfrac{\sigma^2}{n}}} = \dfrac{x_i - \bar{x}}{\sqrt{\sigma^2 + \dfrac{\sigma^2}{n}}}$ (A.2)

Since the Student's t distribution is defined by

$$\frac{Z}{\sqrt{V/r}} \sim t_{(r)} \tag{A.3}$$

where $Z \sim N(0,1)$ and $V \sim \chi^2_{(r)}$, then from (A.2) and (A.3)

$$\frac{\dfrac{x_i - \bar{x}}{\sqrt{\sigma^2 + \dfrac{\sigma^2}{n}}}}{\left[\dfrac{\dfrac{r\hat{s}^2}{\sigma^2}}{r}\right]^{\frac{1}{2}}} \sim t_{(r)} \tag{A.4}$$

Simplifying (A.4) we get

$$\frac{x_i - \bar{x}}{\hat{s}} \sqrt{\frac{r + 1}{r + 2}} \sim t_{(r)}$$

and the random variables

$$T = \frac{Z}{\sqrt{V/r}} = \frac{x_i - \bar{x}}{\hat{s}} \sqrt{\frac{r + 1}{r + 2}} \tag{A.5}$$

Thus for $P(T \geq t) = P(T \geq \dfrac{x_{50} - \bar{x}}{s} \sqrt{\dfrac{r + 1}{r + 2}}) = .02$

The exact 50-year flow for small samples n is

$$\frac{x_{50} - \bar{x}}{\hat{s}} \sqrt{\frac{r + 1}{r + 2}} = t_{.02,r}$$

$$x_{50} = \bar{x} + t_{.02,r} \sqrt{\frac{r + 2}{r + 1}}\ \hat{s} ,$$

or in general

$$x_d = \bar{x} + t_{d,r} \sqrt{\frac{n + 1}{n}}\ \hat{s} \qquad (A.6)$$

APPENDIX B

TRANSFORMATION OF χ^2

We need to find

$$\mathrm{Var}(X_d) = \left(\frac{h\sigma}{n-1}\right)^2 \mathrm{Var}\left(\frac{\sqrt{n-1}\,S}{\sigma}\right) + \mathrm{Var}(\bar{X}) \qquad \text{(B.1)}$$

which means

1) find $\mathrm{Var}(\bar{X})$

and 2) find $\mathrm{Var}\left(\frac{\sqrt{n-1}\,S}{\sigma}\right)$.

We know $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$, so now we must find

$$\mathrm{Var}\left(\frac{\sqrt{n-1}\,S}{\sigma}\right).$$

Let $\quad \gamma = \left(\frac{\sqrt{n-1}\,S}{\sigma}\right)^2 \sim \chi^2_{(n-1)}$ for $0 < \gamma < \infty$

where

$$f(\gamma) = \frac{1}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}}\,\gamma^{\frac{1}{2}r-1}\,e^{-\frac{1}{2}\gamma} \qquad \text{(B.2)}$$

is the density function of (γ), $\Gamma(r/2)$ is the gamma function, and $r = n-1$.

Now we define the one to one transformation

$$v = \sqrt{\gamma}$$

for

$$0 < v < \infty \quad .$$

Then $\gamma = v^2$, and $d\gamma = 2vdv$.

Now consider

$$P(a < v < b) = P(a^2 < \gamma < b^2) \tag{B.3}$$

$$P(a < v < b) = \int_{a^2}^{b^2} \frac{1}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}} \gamma^{\frac{1}{2}r-1} e^{-\frac{1}{2}\gamma} d\gamma \tag{B.4}$$

$$P(a < v < b) = \int_a^b \frac{1}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}} (v^2)^{\frac{1}{2}r-1} e^{-\frac{1}{2}v^2} 2vdv \quad . \tag{B.5}$$

Therefore

$$g(v) = \frac{2}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}} v^{r-1} e^{-\frac{1}{2}v^2} \quad . \tag{B.6}$$

Now by definition

$$\mathrm{Var}\left(\frac{\sqrt{n-1}\ S}{\sigma}\right) = E\left[\frac{\sqrt{n-1}\ s}{\sigma}\right]^2 - \left[E\left(\frac{\sqrt{n-1}\ s}{\sigma}\right)\right]^2 \tag{B.7}$$

and hence by moment generating functions we can find

$$E\left[\frac{\sqrt{n-1}\ s}{\sigma^2}\right]^2 = M''\ (t=0) \tag{B.8}$$

and

$$E\left(\frac{\sqrt{n-1}\ s}{\sigma}\right)^{2} = M'^{2}\ (t=0) \ . \tag{B.9}$$

Let

$$\Psi(r) = \frac{2}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}} \text{ in (B.6)}$$

and thus

$$M(t) = \psi(r)\int_0^{\infty} v^{r-1}\, e^{-\frac{1}{2}v^2}\, e^{tv}\, dv.$$

Since $\frac{de^{tv}}{dt}$ is continuous between 0 and ∞ , then

$$M'(t) = \psi(r)\int_0^{\infty} v^{r-1}\, e^{-\frac{1}{2}v^2}\, (ve^{tv})dv$$

$$E\left(\frac{\sqrt{n-1}\ S}{\sigma}\right) = M'(0) = \psi(r)\int_0^{\infty} v^{r-1}\, e^{-\frac{1}{2}v^2}\, vdv \tag{B.10}$$

$$M''(t) = \psi(r)\int_0^{\infty} v^{r-1}\, e^{\frac{1}{2}v}\, v(ve^{tv})\ dv$$

$$E\left(\frac{(n-1)^{\frac{1}{2}} s}{\sigma}\right)^{2} = M''(0) = \psi(r)\int_0^{\infty} v^{r+1}\, e^{-\frac{1}{2}v^2}\, dv \tag{B.11}$$

Now we need to get M'(0) and M''(0) in the form

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\, e^{-x}\, dx \tag{B.12}$$

in order to integrate. Let

$$x = \frac{r^2}{2}, \quad \text{and} \quad dx = vdv$$

then (B.10) becomes

$$M'(0) = \psi(r)\int_0^\infty v^{2^{\frac{1}{2}r-\frac{1}{2}}} e^{-\frac{1}{2}v^2} vdv$$

$$= \psi(r)\int_0^\infty (\tfrac{2}{2})^{(\frac{1}{2}r + \frac{1}{2})-1} v^{2^{(\frac{1}{2}r + \frac{1}{2})-1}} e^{-\frac{1}{2}v^2} vdv$$

$$= \psi(r)\int_0^\infty 2^{\frac{1}{2}r - \frac{1}{2}} (\tfrac{r}{2})^{2^{(\frac{1}{2}r + \frac{1}{2})-1}} e^{-\frac{1}{2}v^2} vdv$$

$$= \frac{2}{\Gamma(\frac{r}{2})2^{\frac{1}{2}r}} 2^{\frac{1}{2}r - \frac{1}{2}} \Gamma(\tfrac{1}{2}r + \tfrac{1}{2})$$

Thus, $$M'(0) = \frac{\Gamma(\frac{r}{2} + \frac{1}{2})}{\Gamma(\frac{r}{2})} 2^{\frac{1}{2}} \qquad \text{(B.13)}$$

Computing M''(0) in much the same way, we find

$$M''(0) = r \qquad \text{(B.14)}$$

Thus

$$\mathrm{Var}(\frac{\sqrt{n-1}}{\sigma} S) = r - 2\left[\frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})}\right]^2$$

and (B.1) becomes

$$\mathrm{Var}(x_d) = (\frac{h\sigma}{\sqrt{r}})^2 \left\{ r-2\left[\frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})}\right]^2 \right\} + \frac{\sigma^2}{n} \qquad \text{(B.15)}$$

Simplifying (B.15) and by using (B.6)

$$\mathrm{Var}(x_d) = \frac{\Psi t^2_{d,r} (r + 1) + r}{(r + 1) r} \sigma^2 .$$

APPENDIX C

A LIMIT

We want to show that

$$\lim_{r \to \infty} \left\{ r - 2 \left[\frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})} \right]^2 \right\} = \tfrac{1}{2}$$

Let

$$A(r) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})}$$

Using Stirling's approximation to the gamma function

$$\Gamma(z) \sim e^{-z}\, z^{z-\frac{1}{2}}\, (2\pi)^{\frac{1}{2}}\left[1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51840z^3} - \frac{571}{248820z^4} + \dots\right]$$

for

$$(z \longrightarrow \infty \quad \text{in} \; |\arg z| < \pi),$$

it can be shown that

$$\lim_{r \to \infty} A^2(r) \sim \lim_{r \to \infty} r\left(1 + \frac{1}{r}\right)^r \frac{1}{2e}.$$

Now we expand $r(1 + \frac{1}{r})^r$

$$\lim_{r \to \infty} \left\{ \frac{1}{2e} \; r\left[1 + \frac{r}{r} + \frac{r(r-1)}{2!\, r^2} + \frac{r(r-1)(r-2)}{3!\, r^3} + \dots\right] \right\}.$$

By carrying out the multiplication in the numerator and simplifying each term in the brackets, we get

$$\lim_{r \to \infty} \frac{1}{2e} r[e - \frac{1}{r} (\sum_{n=1}^{\infty} \frac{1}{2(n-1)!})]$$

Thus,

$$\lim_{r \to \infty} A^2(r) \sim \frac{1}{2e} r[e - \frac{1}{2r} e] = \frac{r}{2} - \frac{1}{4} .$$

Hence,

$$\lim_{r \to \infty} [r - 2A^2(r)] \sim \lim_{r \to \infty} [r - 2(\frac{r}{2} - \frac{1}{4})] = \frac{1}{2} .$$

(This derivation is gratefully acknowledged to Dr. Don Davis.)

REFERENCES

Aitchison, J., Choice Against Chance, Addison-Wesley Publishing Co., Reading, Mass., 1970. p. 98-120.

Benson, M. A., Characteristics of Frequency Curves Based on a Theoretical 1000 Year Record, Open file report, U.S. Geol. Surv., 1952.

Biswas, A. K., "Some Thoughts on Estimating Spillway Design Flood," Bulletin of the International Association of Scientific Hydrology, XVI, 4.12, 1971.

Carnahan, B., H. A. Luther, and J. O. Wilkes, Applied Numerical Methods, John Wiley and Sons, New York, 1969. p. 100-102.

Clainos, Deme, personal communication, Department of Systems and Industrial Engineering, The University of Arizona, 1972.

Davis, D. R., C. C. Kisiel, and L. Duckstein, Bayesian decision theory applied to design in hydrology, Water Resources Res. 8(1):33, 1972.

Draper, N. R., and H. Smith, Applied Regression Analysis, John Wiley and Sons, Inc., New York, 1966. p. 122.

Fisher, R. A., The Design of Experiments, Oliver and Boyd, London, England, Fifth edition, 1949. p. 182.

Green, P. E., and D. S. Tull, Research for Marketing Decisions, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970. p. 17-18.

Hardison, C. H., "Accuracy of Streamflow Characteristics," in Geological Survey Research, U. S. Geol. Surv., Professional Paper 750C, 1969, p. C228-236.

Hardison, C. H., and M. E. Jennings, "Bias in Computed Flood Risk," J. of Hydraulics Div., Proceedings, ASCE, H43:415-427, 1972.

Hogg, R. B., and A. T. Craig, Introduction to Mathematical Statistics, Macmillan, London, 1971. p. 164.

Klausner, R. F., Evaluation of risk in marine capital investment, Engr. Econ., 14(4):183-214, 1969.

Matalas, N. C., Optimum Gaging Station Location, Proceedings of IBM Scientific Computing Symposium on Water and Air Resource Management, Thomas J. Watson Research Center, Yorktown Heights, New York, 1967. p. 473-489.

Morrison, D. R., Multivariate Statistical Methods, McGraw-Hill, New York, 1967. p. 150-160.

Nash, J. E., and J. Amorocho, The accuracy of the prediction of floods of high return period, Water Resources Res. 4(6):1361-1369, 1966.

Raiffa, H., and R. Schlaifer, Applied Decision Theory, Harvard University, Cambridge, Mass., 1961. p. 228.

Skelton, J., and A. Homyk, A Proposed Streamflow Data Program for Missouri, Open file report, U. S. Geol. Surv. Water Resources Division, Rolla, Missouri, 1970.

Smillie, K. W., An Introduction to Regression and Correlation, Academic Press, New York, 1966. p. 15-20.

Thomas, D. M., and M. A. Benson, "Generalization of Streamflow Characteristics from Drainage Basin Characteristics," U. S. Geol. Surv. Water Supply Paper 1972, 55, 16 Figures. 1970.

Water Resources Council, Bulletin No. 15, "A Uniform Technique for Determining Flood-Flow Frequencies." Washington, D.C., December, 1967.