INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road. Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9325157

A Bayesian approach to estimating a correlation with missing data

Torres-Quevedo, Rocio, Ph.D.

City University of New York, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A BAYESIAN APPROACH TO ESTIMATING A CORRELATION WITH MISSING DATA

by

ROCIO TORRES-QUEVEDO

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1993

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

2/23/93
Date

Chair of Examining Committee
Alan L. Gross

2/23/93
Date

Executive Officer
Carol Kehr Tittle

Professor David Rindskopf

Professor Carol Kehr Tittle

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

22\approv

Abstract

A BAYESIAN APPROACH TO ESTIMATING

A CORRELATION

WITH MISSING DATA

by

Rocio Torres-Quevedo

Adviser: Professor Alan Gross

A Bayesian approach was developed to estimate ρ (correlation between two continuous variables) for the cases where one of the variables is either missing at random or missing completely at random. A formula for the posterior distribution of ρ was developed, evaluated as a function of different factors (n = sample size, r* = correlation between both variables using only the complete cases, R_1=proportion of missing data, R_2 = the ratio of the variance of X using only the complete cases to the variance of X using all cases), and compared to the Maximum likelihood approach for estimating ρ.

Maximum Likelihood based Confidence Intervals and Bayesian Highest Density Regions were computed for different values of n, r*, R_1 and R_2. The Bayesian approach was most advantageous under the following conditions; the prior distribution although "non-informative", restricted ρ to be

positive, both n and r* were small. For larger sample sizes the Highest Density Regions and the Confidence Intervals were more similar. However, when r* was small the Bayesian approach was still more precise than the Maximum Likelihood approach even for larger samples.

Maximum Likelihood Estimates (MLE's), posterior means and posterior medians were also computed for the different values of n, r*, R_1 and R_2. When ρ was not restricted to be positive the Bayesian estimates and the MLE's were practically equal except for small sample sizes where the posterior mean was smaller than the MLE. When ρ was restricted to be positive, the Bayesian estimates tended to be higher than the MLE's; but as the sample size and the r* values increased, the MLE's and Bayesian estimates became more similar.

Both the posterior mean and median were analyzed in order to obtain descriptive information on how they varied as a function of n, r*, R_1 and R_2. As r* increased the Bayesian estimates of ρ strongly increased. As n increased the estimates decreased. Similarly, as R_1 increased the estimates decreased and as R_2 increased the posterior mean and median decreased.

ACKNOWLEDGEMENTS

This dissertation marks the end of a long and thoroughly enjoyable academic sojourn. I am specially grateful to my parents , Carmen Gonzalez and Francisco Torres-Quevedo , who provided emotional and material support throughout. I want to thank my fiance, Eric Karl, for his encouragement, support, and patience as I worked on my dissertation full-time, worked two part-time jobs and interviewed for full-time jobs.

To my friends and classmates Terry Perlis and Rafael Risemberg, thank you for your emotional support throughout the past four years.

I was very fortunate to have an excellent committee. Alan Gross, my adviser, was willing to meet with me frequently in order to meet my aggressive completion schedule. He was also very brilliant, supportive and patient. Carol Kehr Tittle, my department chair, who always supported me throughout my doctorate program and whom I consider a role model. David Rindskopf, a former professor, whose technical expertise I admire. Also Theodore Abramson and Manuel Martinez-Ponz who graciously agreed to read my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

Chapter I

INTRODUCTION

A frequent problem in educational practice arises when one wants to estimate the correlation (ρ) between two variables X and Y and there is incomplete information or missing data on Y. The way in which the data are missing varies. Firstly, the data can be missing at random with respect to X and Y. This case is referred to as "data missing completely at random (MCAR)". Secondly, the data can be missing as a function of X alone. This case is labeled as "data missing at random (MAR)" and is known as "restriction of range" in the psychometric literature. Thirdly, the data can be missing as a function of Y or as a function of both Y and X. This cases referred to as "data not missing at random (NMAR)".

As an example of data MCAR, consider the case where job applicants are given an aptitude test X and they are all recruited regardless of their X scores. The institution then wants to validate the aptitude test X. A sample of the original applicants who took test X is randomly selected after one year and their job performance (Y) is evaluated. While X scores are observed for all applicants, job performance is observed only for the randomly selected cases. The data are MCAR because the probability that job performance is recorded

is the same for all applicants regardless of their test scores or their job performance. As an example of data MAR consider the case where applicants to a school program are selected according to their scores on test X, that is, only applicants with highest scores on test X will be accepted in the program. The school is then interested in estimating the correlation between test X and academic success. Academic success (Y) is measured as the GPA that the selected students obtain after their first year at the program. Scores on test X will be observed for all applicants, however, the Y scores for those applicants whose test scores were lowest will be missing. In this case the data will be MAR because the probability that Y is observed is a function of X. However, suppose that not all applicants who were selected enter the program; rather, they accept offers at other schools with higher selection standards. These students are likely to be high scorers on unmeasured variables which may be related to Y. Thus, the probability that Y is observed may depend not only on X but also on the potential value of Y. Therefore, in this case the data would be NMAR.

In estimating ρ in cases 1 and 2 using the maximum likelihood methods, one can ignore the missing data process and simply analyze the observed data (the XY scores of the complete cases and the X scores of cases missing Y. In these cases, the MLE is straightforward, due to the simple monotonic

pattern of missing values. However, in other cases, for example in case 3, the estimation must take into account the missing data process. In effect, one must simultaneously estimate the parameters of the data model and parameters of the missing data process. This third situation is referred to as "nonignorable".

This dissertation considers a Bayesian approach to estimating ρ for the MCAR and the MAR cases, and compares this estimate to the Maximum likelihood estimate. Although it would be of interest to consider the NMAR situation, this case would be very complex since one must specify a model for the data missing process as well as a prior distribution for the parameter underlying this process.

Cohen (1955, 1957) introduced the maximum likelihood estimation method (MLE) to estimate ρ when data is MCAR or MAR. The maximum likelihood estimate ρ under bivariate normality represents the traditional formula for the correction for restriction of range

$$\hat{\rho} = \hat{\beta}_1 \hat{\sigma}_x / (\hat{\beta}_1^2 \hat{\sigma}_{xx} + \hat{\sigma}_{yy|x})^{1/2} \qquad [1]$$

$$= r^* / [r^{*2} + R(1-r^{*2})]^{1/2} \qquad [2]$$

Where:

$\hat{\beta}_1$: Maximum likelihood estimate of the regression

weight (slope): sample regression weight using the complete cases

$\hat{\sigma}_{xx}$: Maximum likelihood estimate of the variance of X (sample variance of X using both complete and incomplete cases)

$\hat{\sigma}_{yy|x}$: Maximum likelihood estimate of the residual variance or mean square error in predicting Y from X (sample residual variance using the complete cases)

r^*: The XY correlation for the selected cases (complete cases)

R: The ratio of the variance of X in the selected group relative to the variance of X in the total group.

The standard error (SE) for $\hat{\rho}$ [$SE(\hat{\rho})$] is estimated using standard maximum likelihood approaches. Thus

$$SE(\hat{\rho}) = (\underline{d}' \; C \; \underline{d})^{1/2} \qquad [3]$$

Where:

$\underline{d}$: A vector of the partial derivatives of $\hat{\rho}$ with respect to the parameters (means, variances and covariance) underlying the bivariate normal distribution of X and Y different parameter

estimates

C: A matrix of the variances and covariances for the MLE estimates of the bivariate normal parameters.

Considering that the Maximum likelihood estimators are asymptotically normally distributed, one could test the significance of ρ using a t-test where $t=\hat{\rho}/SE(\hat{\rho})$, or calculate a confidence interval for $\hat{\rho}$. In terms of obtaining an interval estimate of ρ, one could consider a Fisher Z-transform of $\hat{\rho}$, and treat $Z(\hat{\rho})$ as normally distributed with mean $Z(\rho)$ and variance $1/(n_c-3)$, where n_c = number of cases measured on X and Y. It should be noted that the confidence intervals derived in this manner will only be approximate, since the variance of $\hat{\rho}$ is not the same as the variance of the estimator of ρ when there are complete data.

The MLE method is strongly dependent on the large sample properties of $\hat{\rho}$, i.e., that in large samples $\hat{\rho}$ is normal with mean ρ, and approximate standard error $SE(\hat{\rho})$. For small samples $\hat{\rho}$ may not be normal and $SE(\hat{\rho})$ might be smaller than the true standard error. Thus the usefulness of this method in small samples may be problematic. In addition, the MLE method does not allow one to incorporate prior information concerning ρ. For example, very often one can tell a priori if a correlation is positive or negative. The incorporation of this

prior information would increase the accuracy of the estimation.

This study presents an alternative method for estimating ρ for the ignorable cases (i.e., MAR and MCAR cases). We consider a data set where X is observed for n cases, but Y is observed for only $n_c<n$ cases. A Bayesian approach is used to obtain a posterior distribution for ρ when Y is MAR or MCAR. This approach, unlike the MLE method, yields exact probability statements in small samples. In addition, the Bayesian approach allows for the incorporation of prior information. An expression for the posterior distribution of ρ is derived for the ignorable cases and where ρ is limited to have a value within a certain range. Next, this formula is used to obtain highest density or probability intervals for ρ as a function of $R_1 = n_c/n$, n, r^*, $R_2=s_x^2*/s_x^2$; where, r* is the sample correlation using only the complete cases; s_x^2* is the sample variance of X using only the complete cases; and s_x^2 is the sample variance of X using all the cases. The Bayesian probability intervals for ρ are compared with the MLE confidence intervals, which are derived in terms of an adaptation of the Fisher's Z transform to the missing data case. Cases where the Bayesian approach is superior are identified.

Chapter II

REVIEW OF THE LITERATURE

A. Maximum likelihood estimate

The maximum likelihood estimator for ρ was first studied by Cohen (1955, 1957). For XY distributed bivariate normal, estimates for ρ were provided under three different missing data processes. In all three cases the Y scores are MAR. The three cases are as follows: truncation, X and Y are observed for N cases such that $X \leq X_0$; censoring, the same as the previous cases except the number of cases for whom $X \leq X_0$ are noted; and finally, the cases studied in the present research (selected samples) where X is completely observed but Y is MAR. The MLE for ρ in this case is the traditional restriction of range correction formula (equations 1 and 2) for X completely observed and Y only partially observed. These results were generalized to the multivariate case in a later paper (Cohen, 1957).

It should be noted that Cohen's maximum likelihood method estimate for ρ is based on the assumption of bivariate normality and Y MAR. With a modification for the estimate of the variance of X, the Maximum likelihood estimate would also apply under the less stringent assumption that the conditional

distribution of Y given X is normal with a linear and homoscedastic regression and further, Y is MAR. The psychometric literature has considered both the sampling properties of $\hat{\rho}$ and the robustness of $\hat{\rho}$ with respect to the violation of these underlying assumptions. First, we will consider those studies that looked at the sampling properties of $\hat{\rho}$. Next, we will consider papers that dealt with the violation of the distribution assumptions. Next, we will review those studies that considered violations of the MAR assumption. Finally, we will review the papers that examined the violation of both assumptions.

B. Sampling properties of $\hat{\rho}$

Gross and Kagen (1983) compared the expected mean square error (EMSE) for the corrected correlation with the EMSE for the uncorrected correlation. Since EMSE = $BIAS^2$ + SAMPLING VARIANCE, it is possible for one estimator to be more biased than the other one and still have a lower EMSE due to a lower sampling variance. Results revealed that the EMSE value tends to be lower for the uncorrected correlation when ρ is low, selection is extreme and the size of the selected group is small.

Bobko and Rieck (1980) considered the so called delta

method to obtain standard errors for functions of correlation coefficients. An attempt was made to apply this method to obtaining standard errors of $\hat{\rho}$. However, this method is limited since it considers R, the ratio of the variance of X in the selected group relative to the variance of X in the total group to be a constant, when in fact it is a random variable.

Gross and Perry (1983) considered a Bayesian approach for obtaining the probability distribution of the squared XY correlation in a future sample where Y was MAR. Although a Bayesian approach was used, the population correlation was not the parameter of interest.

Tanner and Hung Wong (1987) described an iterative technique based on imputing missing values to obtain the Bayesian posterior distribution of parameters of interest when the available sample contains missing data. This method could be adapted to the current problem of estimating ρ. However, in the present research we will concentrate on a closed form expression for the posterior distribution of ρ.

C. Violation in assumptions

i. Violations of distribution assumptions

Greener and Osburn (1980) examined the effects of varying degrees of violation of the linearity and homoscedastic

assumptions on the accuracy of $\hat{\rho}$. The results indicated that when the linearity assumption and/or the homoscedastic assumption was violated, the MLE of ρ was reasonably accurate for moderate degrees of selection (40% or less screened out). However, for larger degrees of selection, violation of the linearity assumption results in underestimates of the population correlation. In the case where the homogeneity assumption is violated, larger degrees of selection results in underestimates of the population correlations for fan-shaped distributions, and in unacceptable overestimates of the population correlations for football-shaped distributions.

Gross & Fleischman (1987) studied the effect of nonlinearity on the accuracy of the correction for restriction of range formula. Three different XY populations were considered. In one of them the XY relationship was linear. In a second one the regression form was concave and in a third one the regression form was convex. Expected mean square errors were computed for both squared corrected correlations and squared uncorrected correlations (squared correlations between X and Y using only the complete cases). Results indicated that the correction formula was almost as accurate for the concave case as it was for the linear case when the linear XY relationship was relatively high, or the sample size or the proportion selected were relatively high. Indeed, if none of these conditions were present, the correction formula

was less accurate than the uncorrected formula in most cases. In the case of a convex regression form, the uncorrected formula was superior to the corrected one.

Schmidt, Hunter, and Urry (1976) proposed an estimate of ρ that corrects for both restriction of range and unreliability of the Y measure. They proposed a stepwise procedure which involves first correcting both the reliability and validity coefficients for restriction of range as follows

$$r_{yy}' = 1-[(1-r_{yy})/(1-r_{xy}^2\ (1-s_x^{2\prime}/s_x^2))] \qquad [4]$$

$$r^* = r_{xy}(s_x'/s_x)/[1-r_{xy}^2 + r_{xy}^2(s_x^{2\prime}/s_x^2)]^{1/2} \qquad [5]$$

where:

r_{yy}': criterion reliability in the restricted group

r_{yy}: criterion reliability in the unrestricted group

s_x': standard deviation of the predictor in the restricted group

s_x: standard deviation of the predictor in the unrestricted group

r^*: validity in the restricted group

r_{xy}: validity in the unrestricted group

In a third step the validity coefficient already

corrected for restriction of range is corrected for unreliability of the predictor

$$R^{*} = r^{*}/(r_{yy}')^{1/2} \qquad [6]$$

Lee, Miller and Graham (1982) conducted an empirical study in order to evaluate the double correction procedure proposed by Schmidt et al. (1976). The results indicated that the double corrected correlation was a better estimate of ρ than the uncorrected correlation. However, the double corrected method yielded slight overestimates of ρ.

Bobko (1983) analytically examined the properties of the double corrected correlation and showed that it is negatively biased and that this bias is inversely related to sample size and/or selection ratio.

In a paper by Mendoza and Mumford (1987) new double corrected correlation formulas were developed, one for the case where the restriction of range results from direct restriction on X, and another one for the case where restriction of range is the result of direct restriction on a latent variable defined by the predictor.

Gross (1982) demonstrated that the correction for restriction of range formula can yield an accurate estimate of ρ when both the linear and homoscedastic assumptions are

violated. In some cases violations in these two assumptions can be offsetting. This is true when the situation is such that

$$Q = [s_e^2 / s_{es}^2 / (b/b_s)]^{1/2} = 1 \qquad [7]$$

where:

s_e^2: residual variance of the regression of Y on X for the total group.

s_{es}^2: residual variance of the regression of Y on X for the selected group.

b: slope of the regression of Y on X for the total group

b_s: slope of the regression of Y on X for the selected group

In cases where Q>1 the correction for restriction of range formula will yield an overestimate of ρ. For the case where Q<1 the formula will yield an underestimate of ρ.

Holmes (1990) analytically considered the effects of non-linearity and heteroscedasticity on the mathematical representation of ρ in an unrestricted bivariate population, using the population counterpart of the "corrected" correlation computed from the parameters of a selected

population.

ii. Violations of the MAR assumption

The correction for restriction of range formula assumes that Y is MAR, that is, Y is missing only as a function of X. Linn (1968) analyzes the case where there is an additional selection variable, Z, a variable that is not available. He indicates that (a) when the population correlation between X and Z is relatively large, the correction formula would underestimate the correlation between X and Y, (b) the amount of negative bias increases as R increases (ratio between the variance of Z for the total number of cases observed and the variance of Z for the selected group), and (c) when the correlation between Z and X is 1 the correction formula would be an accurate estimate.

Gross and McGanney (1987) illustrate how the MLE method yields biased estimates for the parameters of the regression of Y on X when the selection process is nonignorable. Most typically, the slope coefficient and the residual variance would be underestimated and the regression intercept would be overestimated. The negative biases in the estimation of the slope and residual variance would result in an underestimation of the corrected correlation. Gross and McGanney (1987) also described a statistical method of analysis that yielded better estimates of the XY correlation as well as the regression parameters when the selection process is nonignorable.

Roe and Elshout (1972) considered the case where all the variables on which selection is based are known. They present the following formula to correct for restriction of range under these circumstances

$$R_{yz}'=[(r_{yz}-r_{xy}\ r_{xz})/H_y+R_{xy}\ r_{xz}\ H_x]/[1+r_{xz}^2\ (H_x^2-1)]^{1/2} \quad [8]$$

where:

x:weighted sum of the variables on which selection is based

y:the predictor

z:the criterion

r:correlation in the selected group

R:correlation in the unselected group

H:standard deviation in the unselected group divided by the standard deviation in the selected group

In a later study, Roe (1979) compared the above formula with a similar one in which X is substituted by a variable based on the multiple regression equation resulting from regressing the different selector variables on the probability of being selected. He discussed that in many cases "actual selection" is different from "intended selection" and that

using a multiple regression analysis to create the X variable will yield more accurate results than using the weighted sum.

Several authors (Cohen 1959; Alexander, Alliger, & Hanges, 1984; Dobson 1988) have discussed the case where Y is MAR but X is NMAR due to truncation. In this case unbiased maximum likelihood estimates for the parameters for the regression of Y on X can still be obtained using the complete cases. However, obtaining an unbiased estimator of the variance of X (σ_{xx}) is not as straightforward as it is in the case where X is not missing or is MAR. Cohen (1959) developed the maximum likelihood estimate of the variance of X from a truncated sample. Alexander, Alliger & Hanges (1984) used Cohen's MLE of the variance of X to estimate the population variance of X in order to correct for range restriction. Dobson (1988) presents a method to estimate R (the ratio of the population and sample standard deviations).

The case where both X and Y are truncated has also received attention in the psychometric literature. Thorndike (1947) and Wells & Fruchter (1970) developed different corrected formulae for this case (formula for ρ when there is restriction of range and both variables are truncated). Alexander, Carson, Alliger & Barret (1984) evaluated the accuracy of these two formulae. Alexander et al. (1984) concluded that the Wells and Fruchter method is supperior to

that of Thorndike and that it is quite accurate under most conditions but that it often yields slight overcorrections. Alexander, Carson, Alliger and Carr (1987) presented and evaluated the performance of an improved corrected formula. Alexander, Hanges, and Alliger (1985) extended Cohen's rationale to estimate the population variances of both X and Y when both variables are truncated and substituted these values into the Alexander, Carson, Alliger & Carr (1984) improved corrected formula.

The general problem of estimating the association between two or more variables under nonignorable missing data processes is discussed by Alexander, Barret, Alliger, and Carson (1986).

iii. Violations of both, distribution assumptions and the MAR assumption

Gross and Fleischman (1983) studied the accuracy of the MLE of ρ when both the distribution assumptions and the MAR assumptions are violated. Specifically, the accuracy of the MLE of ρ was studied when: (1) selection on Y is not based on X alone and the regression of Y on X is not linear, and (2) when selection on Y is not based on X alone and it is not homoscedastic either. Results indicated that the MLE of ρ for the complete data set (corrected correlation) was very inaccurate in both cases except when the degree of selection was very small. However, the corrected correlation was a

better estimate of ρ than the uncorrected correlation (MLE of ρ using only the complete cases) under certain conditions. Gross and Fleischman (1983) concluded that when Q>1 and the selection process is not based on X alone the correction formula should not be applied.

Chapter III

METHOD

A. Derivation of the Posterior Distribution of ρ given Missing Data on Y

Consider the general Bayesian approach to obtaining the posterior distribution of some parameter vector $\underline{\theta}$. Let $p'(\underline{\theta})$ represent the prior distribution for $\underline{\theta}$. Given a sample of data (d), the likelihood function is denoted as $L(\underline{\theta}|d)$. The posterior distribution for $\underline{\theta}$ [$p''(\underline{\theta}|d)$] is then proportional to the product of the prior and the likelihood,

$$p''(\underline{\theta}|d) \propto p'(\underline{\theta})\ L(\underline{\theta}|d) \qquad [9]$$

In the present problem we take $\underline{\theta}$ to initially be the parameters of the joint bivariate normal distribution of X and Y,

$$\underline{\Theta}=[\mu_x,\ \mu_y,\ \sigma_{xx},\ \sigma_{yy},\ \rho] \qquad [10]$$

where:

μ_x: mean of X

σ_{xx}: variance of X

σ_{yy}: variance of Y

μ_y: mean of Y

ρ: correlation between X and Y

Following Lindley (1965) a general non informative prior for $\underline{\Theta}$ is given as

$$p'(\underline{\Theta}) \propto (1/\sigma_{xx})(1/\sigma_{yy})\ P'(\rho), \quad [11]$$

where $P'(\rho)$ is the prior distribution for ρ. In the case where there are not missing data a commonly used choice for $P'(\rho)$ is $1/(1-\rho^2)^{3/2}$. Given this choice for $P'(\rho)$, the non informative prior given by (11), is the standard non informative prior for the parameters of a bivariate normal distribution.

In terms of $\underline{\theta}$ the likelihood function for the incomplete data is cumbersome to use. More specifically, the likelihood based on $\underline{\theta}$ is the product of two factors which are conditioned on non-distinct or overlapping parameter sets. For the n_c cases measured on both x and y the likelihood function is conditioned on all five parameters. However, for the n-nc cases measured only on x, the likelihood function is conditioned on a subset of these

parameters, μ_x and σ_{xx}. The construction of the posterior distribution of ρ is far simpler if the likelihood can be factored into two terms, each conditioned on distinct parameters. This can be accomplished if $\underline{\theta}$ is transformed to $\underline{\theta}^*$:

$$\underline{\theta}^*=[\mu_x,\ \gamma_x=1/\sigma_{xx},\ \gamma_y=1/\sigma_{yy|x},\ \beta_0,\ \beta_1] \qquad [12]$$

Then

$$p'(\underline{\theta}^*) \propto (1/\gamma_y^{1/2})\ \ (1/D^{3/2})\ \ P'(\beta_1\ \gamma_y^{1/2}/D^{1/2}) \qquad [13]$$

Where:

$$D = \gamma_x + \beta_1^2\ \gamma_y \qquad [14]$$

We use this prior distribution for $\underline{\theta}^*$ in obtaining the posterior distribution of ρ for the missing data case.

If we have a data set consisting of n X's and nc Y's, then, the likelihood function can be factored as the product of two likelihoods:

$$L(\mu_x,\ \gamma_x)\ L(\beta_0,\ \beta_1,\ \gamma_y) \qquad [15]$$

Where:

$$L(\mu_x, \gamma_x) = (\gamma_x)^{n/2} \exp\{-\gamma_x SS_x/2\} \exp\{-\gamma_x n(\bar{x}-\mu_x)^2/2\} \quad [16]$$

$$L(\beta_0, \beta_1, \gamma_y) =$$

$$(\gamma_y)^{nc/2} \exp\{-\gamma_y SS^*_c/2\} \exp\{-\tfrac{1}{2}(\underline{\beta}-\hat{\underline{\beta}}^*)'(X^{*\prime}X^* \gamma_y)(\underline{\beta}-\hat{\underline{\beta}}^*) \quad [17]$$

SS_x: sum of squares of X computed from n X scores in the total sample

$\bar{x}$: mean of X computed from n X scores

SS_c^*: error sum of squares computed predicting Y from X using the nc XY scores

$\hat{\underline{\beta}}^*$: regression weights computed using the nc XY scores

X^*: n_c by 2 matrix for the complete sample X scores, where the first column is a vector of unities and the second column is a vector of the n_c X scores

Multiplying the prior for $\underline{\theta}^*$ given in [13] and the likelihood function for $\underline{\theta}^*$ given in [15] yields an expression proportional to the joint posterior for $\underline{\theta}^*$. If we integrate over μ_x and β_0, we get the joint posterior distribution for β_1, γ_x, γ_y

$$p''(\gamma_x, \beta_1, \gamma_y) \propto \int_{\mu_x} \int_{\beta_0} L(\mu_x, \gamma_x)\ L(\beta_0, \beta_1, \gamma_y)\ P'(\underline{\theta}^*) \quad [18]$$

$$p''(\gamma_x,\ \beta_1,\ \gamma_y) \propto \{\gamma_x^{(n-1)/2} \exp(-\gamma_x\ SS_x/2)\}$$
$$\{\gamma_y^{(nc-2)/2} \exp(-\gamma_y\ SS^*_c/2)\}$$
$$\{\exp[(-\gamma_y\ SS^*_x/2)(\beta_1-\hat{\beta}^*_1)^2]\}$$
$$(1/D)^{3/2}\quad P'(\beta_1\ \gamma_y^{1/2}/D^{1/2}) \quad [19]$$

To reintroduce the parameter of greatest interest (ρ) we make the following change of variables to

$$\rho,\ V_1=\sigma_{yy}/\sigma_{xx},\ V_2=1/\sigma_{xx} \quad [20]$$

We obtain

$$p''(\rho,\ V_1,\ V_2) \propto V_2^{\alpha-1} \exp(-V_2\beta)$$
$$(1/V_1)^{(nc+1)/2}\ (1/1-\rho^2)^{(nc-1)/2}\quad P'(\rho) \quad [21]$$

Where:

$$\alpha-1 = (n+n_c-4)/2 \quad [22]$$

$$\beta=\{SS_x/[2(1-\rho^2)]\}\{[1-A+BC-2DC^{1/2}\} \quad [23]$$

A: $\rho^2\ (1-R)$

B: R/V_1

C: SS^*_y/SS^*_x

D: $Rr^*\rho/V_1^{1/2}$

R: SS_x^*/SS_x

r^*: correlation between X and Y using the nc XY scores

SS_y^*: sum of squares of Y computed from nc Y scores

SS_x^*: sum of squares of X computed from nc X scores

Integrating over V_2

$$p''(\rho, V_1) \propto [(1/V_1)^w(1/1-\rho^2)^q p'(\rho)]/\beta^\alpha \quad [24]$$

Where:

$$W: (n_c+1)/2 \quad [25]$$

$$q: (n_c-1)/2 \quad [26]$$

$$\alpha: (n+n_c-2)/2 \quad [27]$$

Making a change of variables to ρ, $\Psi=V_1^{1/2}(SS^*_x/SS^*_y)^{1/2}$, we obtain the following final expression

$$p''(\rho, \Psi) \propto [(1-\rho^2)^{(n-1)/2} P'(\rho)]/A^\alpha \quad [28]$$

Where:

$$A: \Psi^{(nc-n+2)/2}\{\Psi[1-\rho^2(1-R)]+(R/\Psi)-(2Rr^*\rho)\} \quad [29]$$

$$\alpha=(n+n_c-2)/2 \quad [30]$$

If we replace R by R_1R_2, where:

$$R_1=n_c/n \qquad [31]$$

$$R_2=s_x^{*2}/s_x^2 \qquad [32]$$

s_x^{*2}: variance of the nc X scores

s_x^2: variance of the n X scores

we can write [28] as

$$p''(\rho,\ \Psi) \propto (1-\rho^2)^{(n-1)/2}\ P'(\rho)/B \qquad [33]$$

where:

$$B:\Psi^{(nc-n+2)/2}\ \{\Psi[1-\rho^2(1-R_1R_2)]+(R_1R_2/\Psi)(2R_1R_2r^*\rho)\}^{\alpha} \qquad [34]$$

If we set nc=n and $SS_x^*=SS_x$, then we consider the case of no missing data and [28] becomes

$$p''(\rho,\ \Psi) \propto (1-\rho^2)^{(n-1)/2}P'(\rho)\ /\ \Psi\ [\Psi+(1/\Psi)-2\rho r^*]^{n-1} \qquad [35]$$

Further, if we set $P'(\rho) = (1/1-\rho^2)^{3/2}$ we obtain

$$P''\ (\rho,\ \Psi) \propto (1-\rho^2)^{(n-4)/2}\ /\ \Psi[\Psi+(1/\Psi)-2\rho r^*]n-1 \qquad [36]$$

which is the standard posterior for ρ given by Lee (1989,

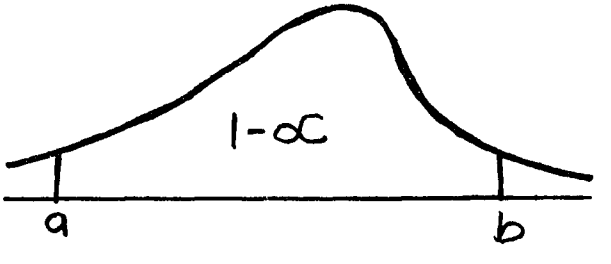
p.171).

B. Calculation of Highest Density Regions, Means and Medians for the Posterior Distribution of ρ

i. Highest Density Region

The Highest Density Region (HDR) of the posterior distribution of a parameter is that interval which is the shortest possible interval for a given probability level ($1-\alpha$). The HDR is also referred to as the Highest Posterior Region, the Credible Interval, and the Bayesian Confidence Interval.

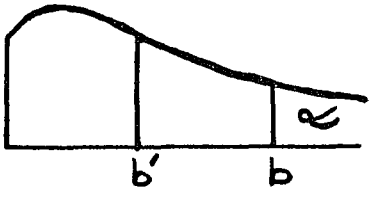
The method used in obtaining the HDR varies as a function of the shape of the posterior distribution of ρ. Three different cases were considered in the present study. In case 1 the interval is found as that region for which the density is equal at either point and which contains a probability of $1-\alpha$. For example, consider the following unimodal distribution of ρ where the density is a continuous function:



Suppose that between a and b lies the desired

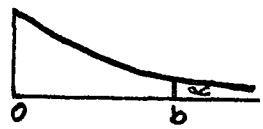
probability $1-\alpha$ and suppose that the density is equal at a and b, then it must be true that for this distribution any point inside the interval has a larger density than any point outside it. Thus [a, b] is the HDR.

This method however cannot always be used because it is sometimes impossible to find an interval that satisfies the equal density condition and that also contains the desired probability. This type of problem arises in two ways when the posterior distribution of ρ is truncated due to the prior distribution. In these cases the density is not a continuous function. For example, suppose that ρ is restricted to be non-negative and that the posterior appears as follows:



The widest interval satisfying the equal density condition [0,b'] may be too narrow to satisfy the probability condition. In this case (case 2) the HDR would be the interval from 0 to b which covers a probability of $1-\alpha$.

Case 3 also arises for a truncated distribution. In this situation, the mode appears at the left lower end point for ρ.



Again the HDR interval is chosen to be [0,b] where this interval contains a probability $1-\alpha$.

ii. Posterior Mean and Median

Both the posterior mean and the posterior median provide Bayesian estimates of ρ. If one considers that estimator of ρ which minimizes the quadratic loss function,

$$E''(\rho-\hat{\rho})^2 = \int (\rho-\hat{\rho})^2 \, p''(\rho) \, d\rho \qquad [37]$$

the loss is minimized for $\hat{\rho}$ = the posterior mean.

Similarly, if one chooses ρ to be the median of the distribution then ρ would be that estimator that minimizes the absolute loss function

$$E'' \, |\rho-\hat{\rho}| = \int |\rho-\hat{\rho}| \, p''(\rho) \, d\rho \qquad [38]$$

Where:

$$\hat{\rho} = \text{Median} \qquad [39]$$

A FORTRAN computer program was written to calculate the 90% HDR's, the means and the medians of the posterior distribution of ρ. These characteristics of the posterior distribution were investigated as a function of the choice for $P'(\rho)$ and the values for n, R_1, R_2, and r^*, where:

$P'(\rho)=(1/1-\rho^2)^{3/2}$; the standard non informative prior.

n=sample size.

$R_1=n_c/n$; proportion of complete data.

$R_2=s_x^{*2}/s_x^2$; ratio of the sample variance of x using the complete cases and the sample variance of x using both the complete and the incomplete cases.

Two different choices were taken for $P'(\rho)$: (a) $P'(\rho)=[1/(1-\rho^2)]^{3/2}$ where $-1 \leq \rho \leq 1$, and (b) $P'(\rho)=(1/1-\rho^2)^{3/2}$ where $0 \leq \rho \leq 1$. In case b, ρ is restricted to be positive.

In addition, n, R_1, R_2 and r^* were assigned the following values:

R_1: .3, .5, .8 (.75 for n=12)

R_2: .5, .75, 1

r^*: 0, .3, .5, .75

n: 12, 20, 30

Although for n=12, R_1=.3 yields a noninteger sample size, for purposes of symmetry with the other n values this R_1 value was not altered.

Considering all different combinations of R_1, R_2, n, and r^* we generated 216 posterior distributions. Using the FORTRAN program we calculated the HDR, mean and median of each of these 216 conditions.

It should be noted that R_1 (the proportion of complete cases) places a restriction on the variance ratio $R_2=s_x^{*2}/s_x^2$. Using expressions for the variance of truncated normal distributions (Maddala, 1983), it can be shown that all of the R_1, R_2 combinations we have used are possible outcomes.

C. Comparison of the Bayesian Estimates with the Maximum Likelihood Estimates

In addition to the 90% HDR's, posterior means and posterior medians, the maximum likelihood estimates (MLE) for ρ and the associated confidence intervals (CI) were calculated for each of the 216 conditions. The confidence intervals were compared to the HDR's and the MLE's were compared to the posterior means and medians.

To obtain the traditional CI the Fisher's Z [$Z(\hat{\rho})$] was applied. For the complete data case where $\hat{\rho} \sim N [\rho, (1-$

$\rho^2)/n)^2]$,

$$Z(\hat{\rho})=1/2\ \ln[(1+\hat{\rho})/(1-\hat{\rho})]\ \sim\ N[Z(\rho),\ 1/(n-3)] \qquad [40]$$

Thus the 90% CI for $Z(\hat{\rho})$, given complete data, would be:

$$Z(\hat{\rho})-1.645\ [1/\sqrt{(n-3)}]\leq 1/2\ \ln[(1+\rho)/(1-\rho)]$$
$$\leq Z(\hat{\rho})+1.645\ [1/\sqrt{(n-3)}] \qquad [41]$$

Solving for ρ

$$(e^A-1)/(e^A+1)\ \leq\rho\leq\ (e^B-1)/e^B+1) \qquad [42]$$

Where:

$$B:\ 2\{Z(\hat{\rho})+1.645\ [1/\sqrt{(n-3)}]\} \qquad [43]$$

$$A:\ 2\{Z(\hat{\rho})-1.645\ [1/\sqrt{(n-3)}]\} \qquad [44]$$

These formulae given in [42], [43], and [44] were modified to obtain 90% CI for the missing data case. The estimate for ρ $(\hat{\rho})$ is given by the MLE, equations [1] and [2]. Further, the sample size was taken to be the number of complete cases (n_c).

Chapter IV

RESULTS

A. Comparison of HDR's and CI's Intervals

Figures 1, 2, and 3 show a comparison of the average widths of the 90% confidence intervals, the 90% HDR's for $-1 \leq \rho \leq 1$, and the 90% HDR's for $0 \leq \rho \leq 1$. Given a specific sample size (n=12, 20, 30) and a specific value for the observed correlation for the complete cases (r^{*}=0, .3, .5, .75), nine 90% confidence intervals were obtained by varying the R_1 (nc/n) and R_2 (s_x^{*2}/s_x^2) conditions. The mean upper and lower bounds were then computed by averaging over the R_1 and R_2 values. For each sample size (n), four 90% averaged CI's were obtained corresponding to the different r* (0, .3, .5, .75) values. A similar process was used to obtain the average 90% HDR's for the cases where $-1 \leq \rho \leq 1$, and $0 \leq \rho \leq 1$. In figure 1, the total sample size is n=12. The length of the lines represent the length of the intervals. The top three lines correspond to the two 90% HDR's ($-1 \leq \rho \leq 1$, $0 \leq \rho \leq 1$) and the 90% CI respectively for r*=0 averaging over all R_1 and R_2 values. Similarly, the three next lines represent the two 90% HDR's ($-1 \leq \rho \leq 1$, $0 \leq \rho \leq 1$) and the 90% CI respectively when r*=.3 averaging all values of R_1 and R_2. The third and fourth sets of three lines pertain to

r*=.5 and r*=.75 respectively. For example, in figure 1 (n=12), given an observed correlation of .3, the Bayesian interval estimate on average ranges from -.386 to .880 when $-1 \leq \rho \leq 1$ and from .049 to .816 when $0 \leq \rho \leq 1$. The Maximum Likelihood Estimate in this case ranges from -.587 to .875. Figures 2 and 3 show the results for n=20 and n=30 respectively.

The major conclusions that can be drawn from these results when ρ is <u>not</u> restricted ($-1 \leq \rho \leq 1$) are the following:

a. In general, the HDR's and the CI's are practically the same when the prior distribution for ρ is non informative on the range $-1 \leq \rho \leq 1$. The upper and lower bounds of the CI's and HDR's are quite similar.

b. When n is small (n=12) both the CI's and the HDR's are, in general, so large that the estimates are most likely of no practical value. For example, in figure 1 (n=12) for r*<.75, all intervals range from a negative value to a high positive value.

c. Even as n increases (n=20, figure 2 and n=30, figure 3), the estimates are very wide for r* ≤.30.

d. In summary, both the Bayesian HDR's and the CI's will not be precise for small n. Even for larger n values, the intervals will be wide for small r* values.

The major conclusions that can be drawn from these results when ρ is restricted to be positive ($0 \leq \rho \leq 1$) are the following:

FIGURE 1
90% HDR AND 90% CI AS A FUNCTION OF r* (n=12)

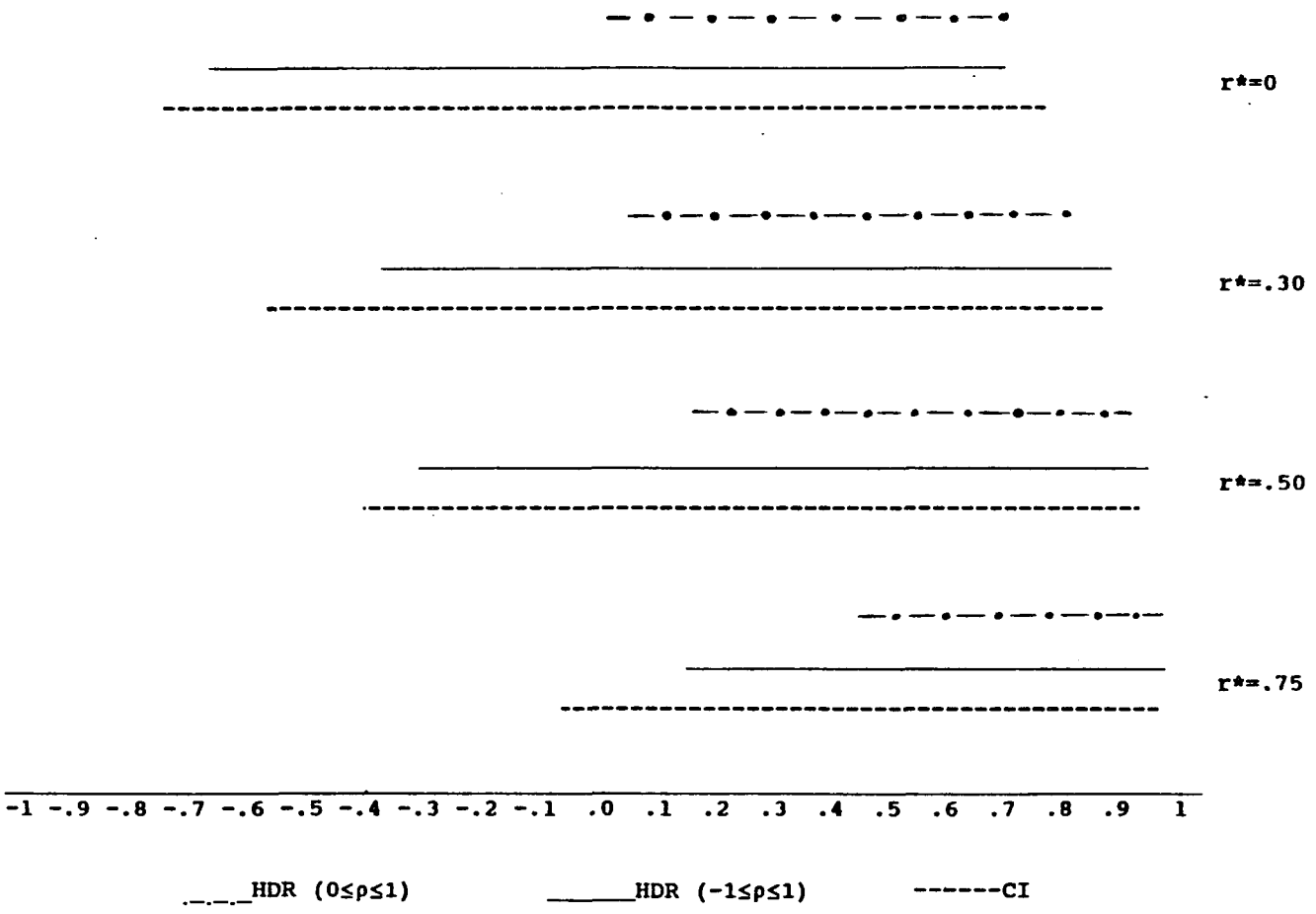


FIGURE 2
90% HDR AND 90% CI AS A FUNCTION OF r* (n=20)

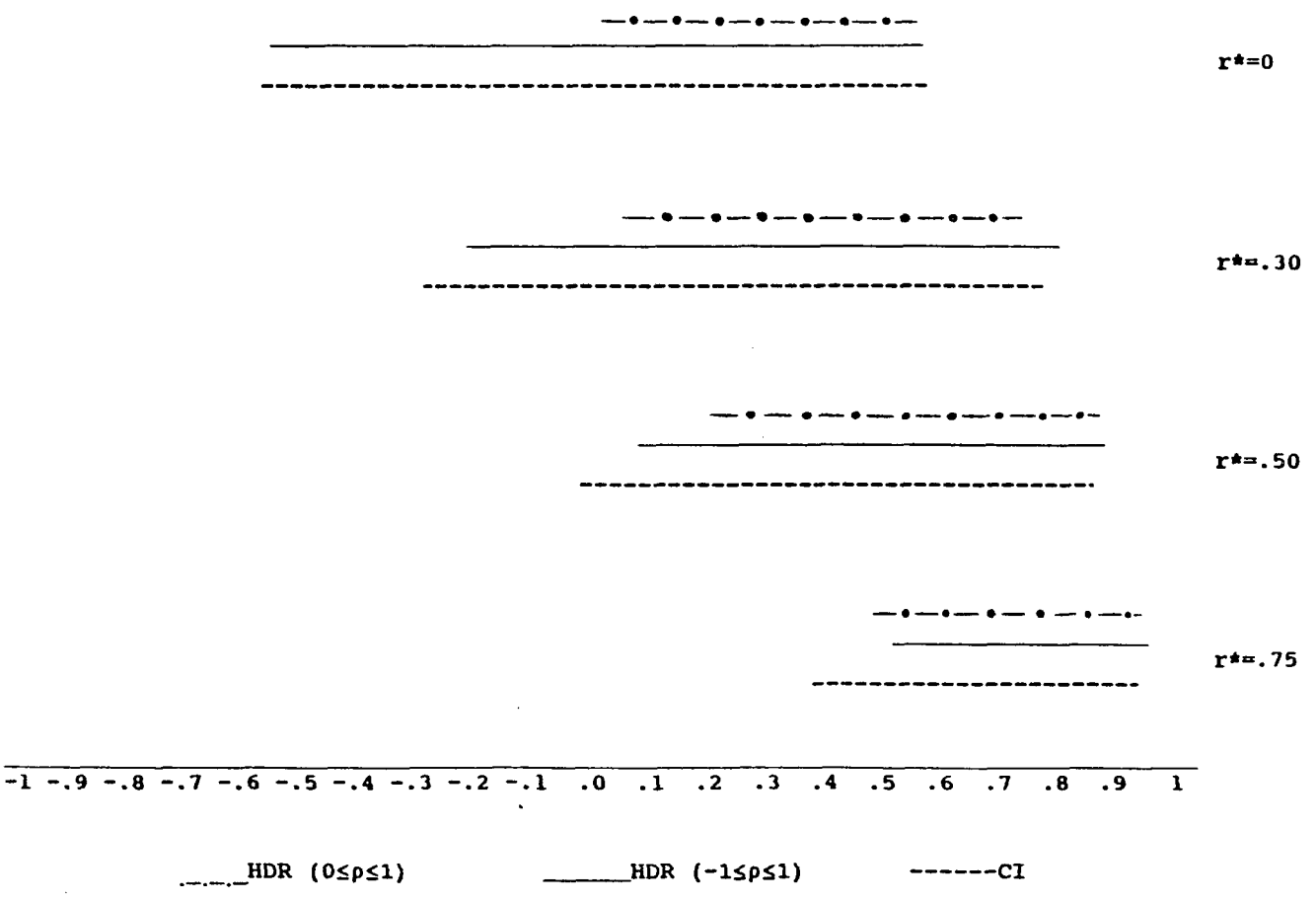
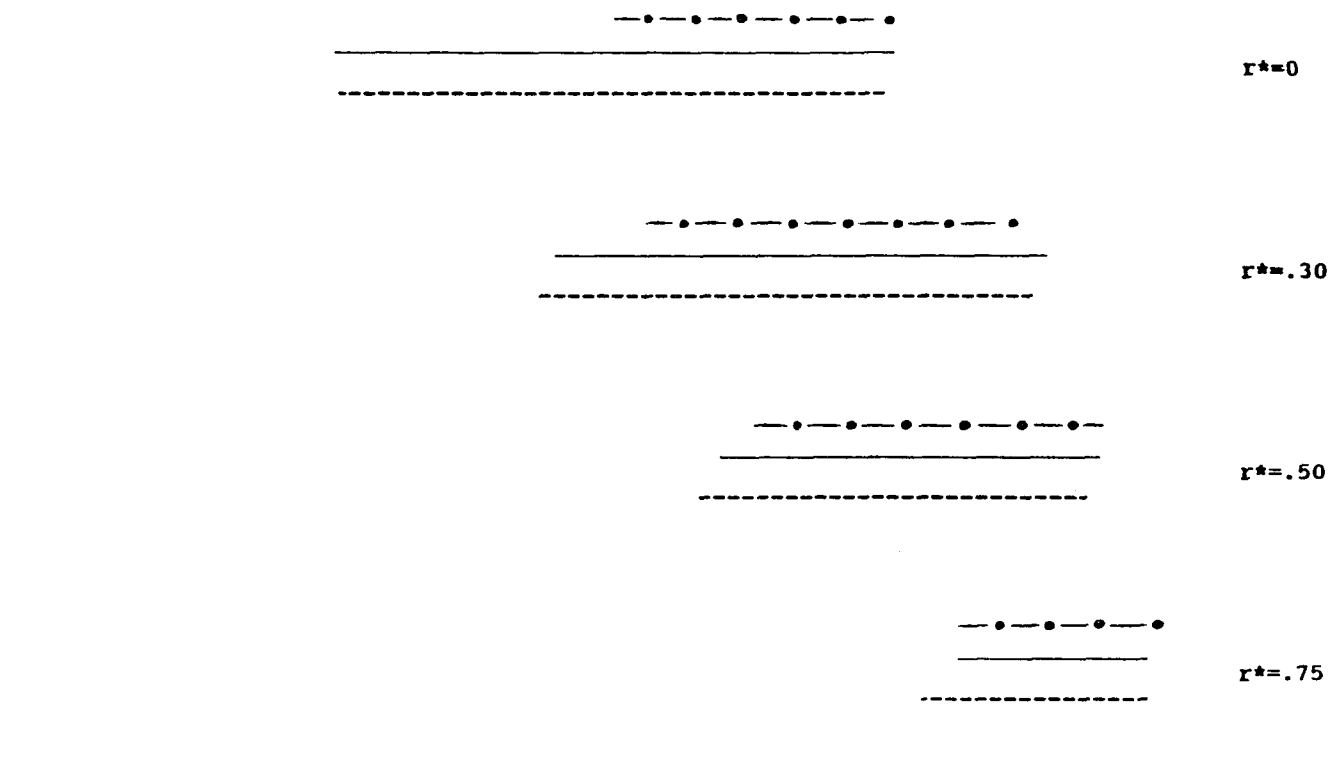


FIGURE 3
90% HDR AND 90% CI AS A FUNCTION OF r* (n=30)



a. Unlike the previous case ($-1 \leq \rho \leq 1$) when ρ is restricted to be positive ($0 \leq \rho \leq 1$) it is evident that the HDR's are narrower and that the bounds are higher than those for the CI's. For example, in figure 2 (n=20), given an observed correlation of .50, the Bayesian estimate on average, ranges from .202 to .866, while the Confidence Interval ranges from -.033 to .856.

b. As the sample size increases the difference between the CI's and the Bayesian intervals decreases. For example, in figure 1 (n=12) given an observed correlation (r*) of .50 the HDR ranges from .164 to .911 and the CI from -.410 to.924, whereas in figure 3 (n=30) for r*=.50 the HDR ranges from .238 to .828 and the CI from .134 to .809.

c. Finally as r* increases the difference between the CI's and the Bayesian intervals decreases. For example, in figure 3 (n=30) for r*=0 the HDR ranges from 0 to .477 and the CI from -.457 to .457. On the other hand, for n=30 and r*=.75 the HDR ranges from .594 to .929 and the CI from .526 to .918.

B. Comparison between the MLE and the Posterior Mean and Median

Figures 4, 5 and 6 show a comparison of the average

values of the MLE's, posterior means and medians. Given a specific sample size (n=12, 20, 30) and a specific value for the observed correlation (r*: 0, .3, .5, .75), nine MLE's, nine posterior means and nine posterior medians were computed by varying the R_1 and R_2 conditions. The average MLE's and the average posterior means and medians were then computed by averaging over all nine values of R_1 and R_2. Thus, for each sample size we obtained four MLE's (one for each value of r*), four posterior means (one for each value of r*) and four posterior medians (one for each value of r*).

In figures 4, 5 and 6 the prior distribution is non-informative on the range $-1 \leq \rho \leq 1$. In figure 4, the total sample size is n=12. Each of the three lines corresponds to a different estimate of ρ ,i.e. the MLE, the posterior mean and the posterior median. As previously, each point is an average of the estimate over all values of R_1 and R_2 for a given value of r*. For example, in figure 4 (n=12), the X axis represent the different values of r* (0, .3, .5, .75) and the Y axis the different values of the estimate of ρ, the MLE, the posterior mean and median. Thus, for n=12 and r*=.3 the averaged posterior mean would be equal to. 263, the MLE would be .349 and the posterior median would be .335. Figures 5 and 6 present the results for n=20 and n=30 respectively. Table 1 presents he values on which figures 4, 5 and 6 are based.

FIGURE 4
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=12, $-1 \leq \rho \leq 1$)

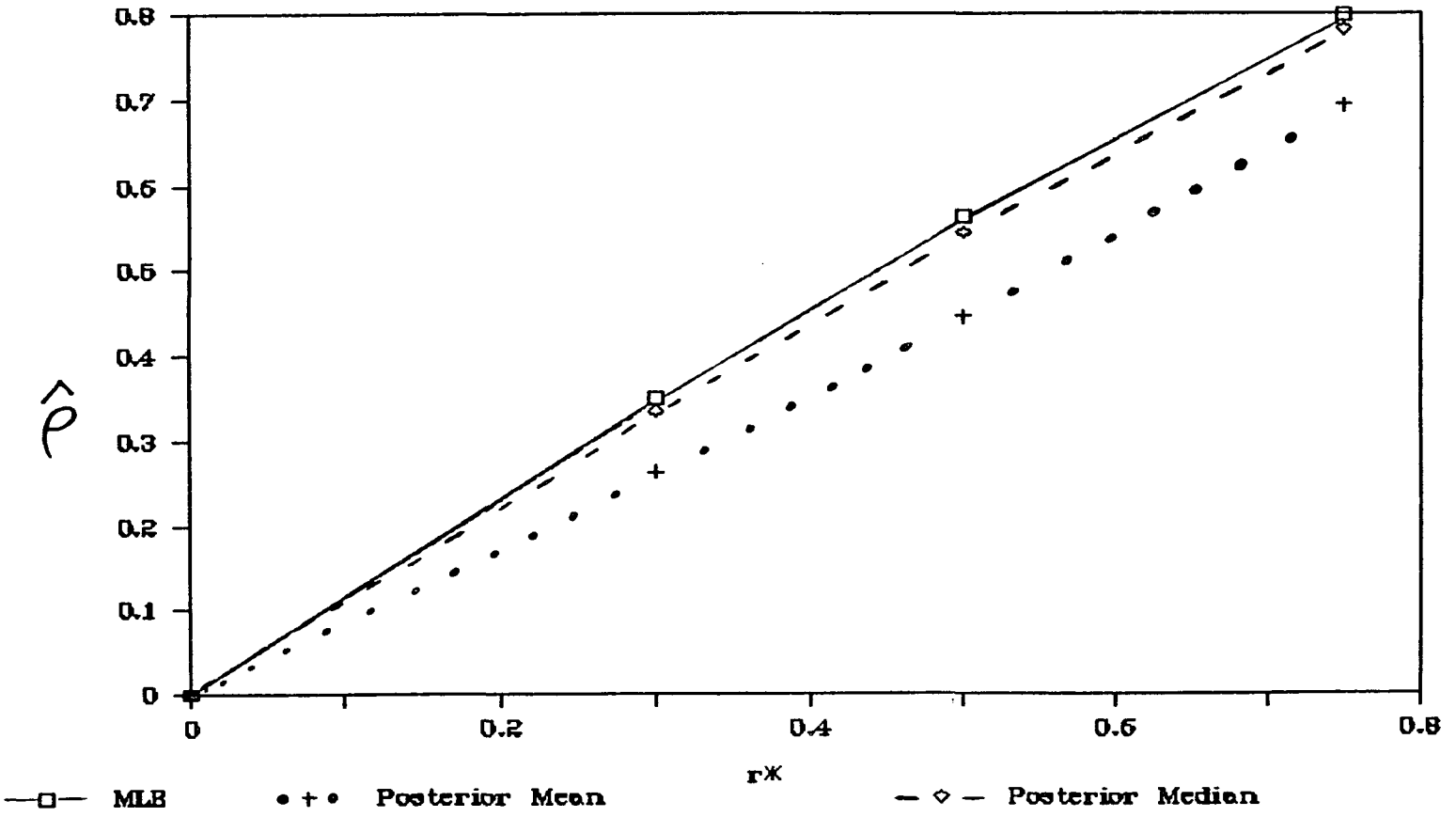


FIGURE 5
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=20 $-1 \leq \rho \leq 1$)

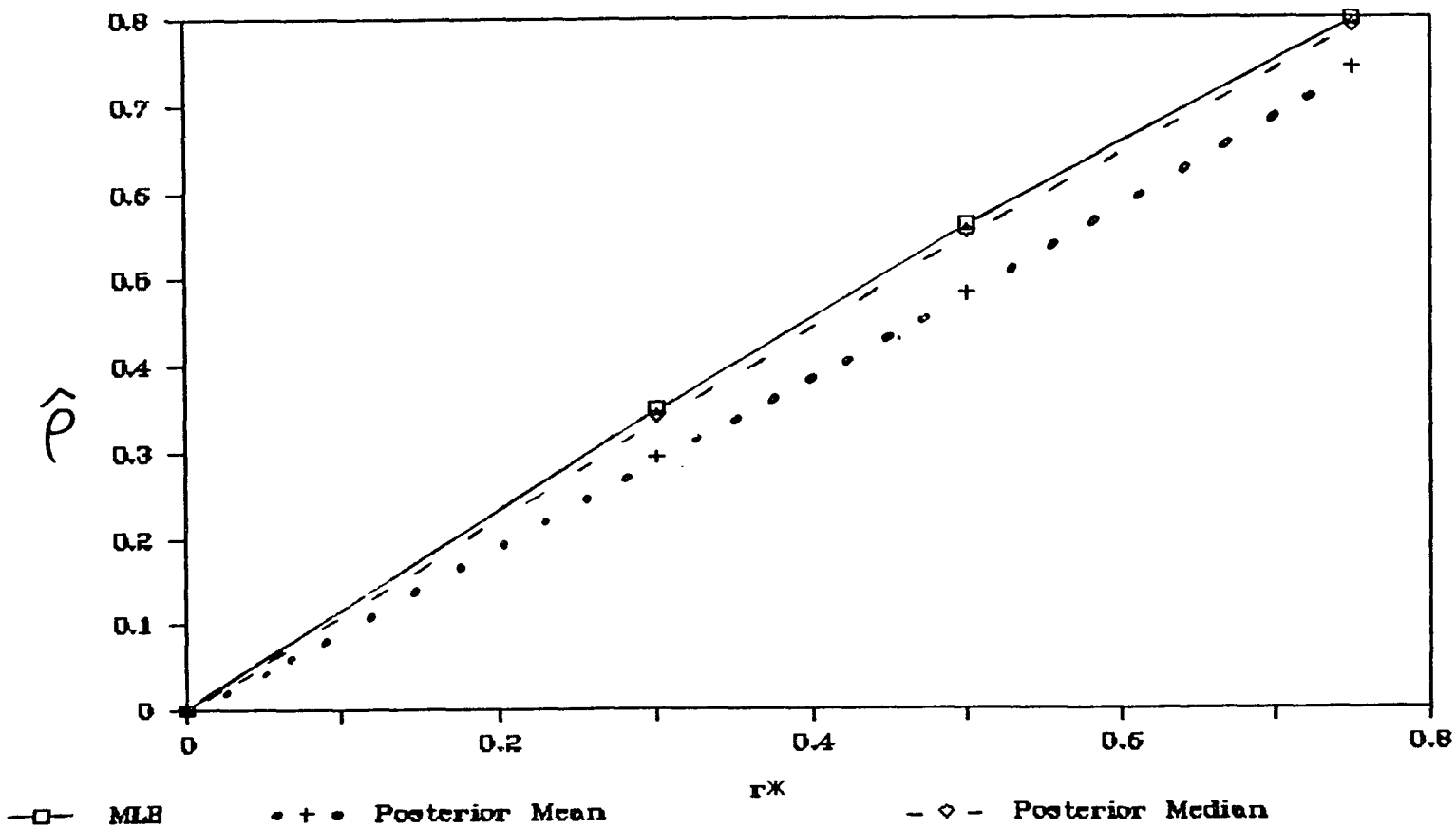


FIGURE 6
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=30 $-1 \leq \rho \leq 1$)

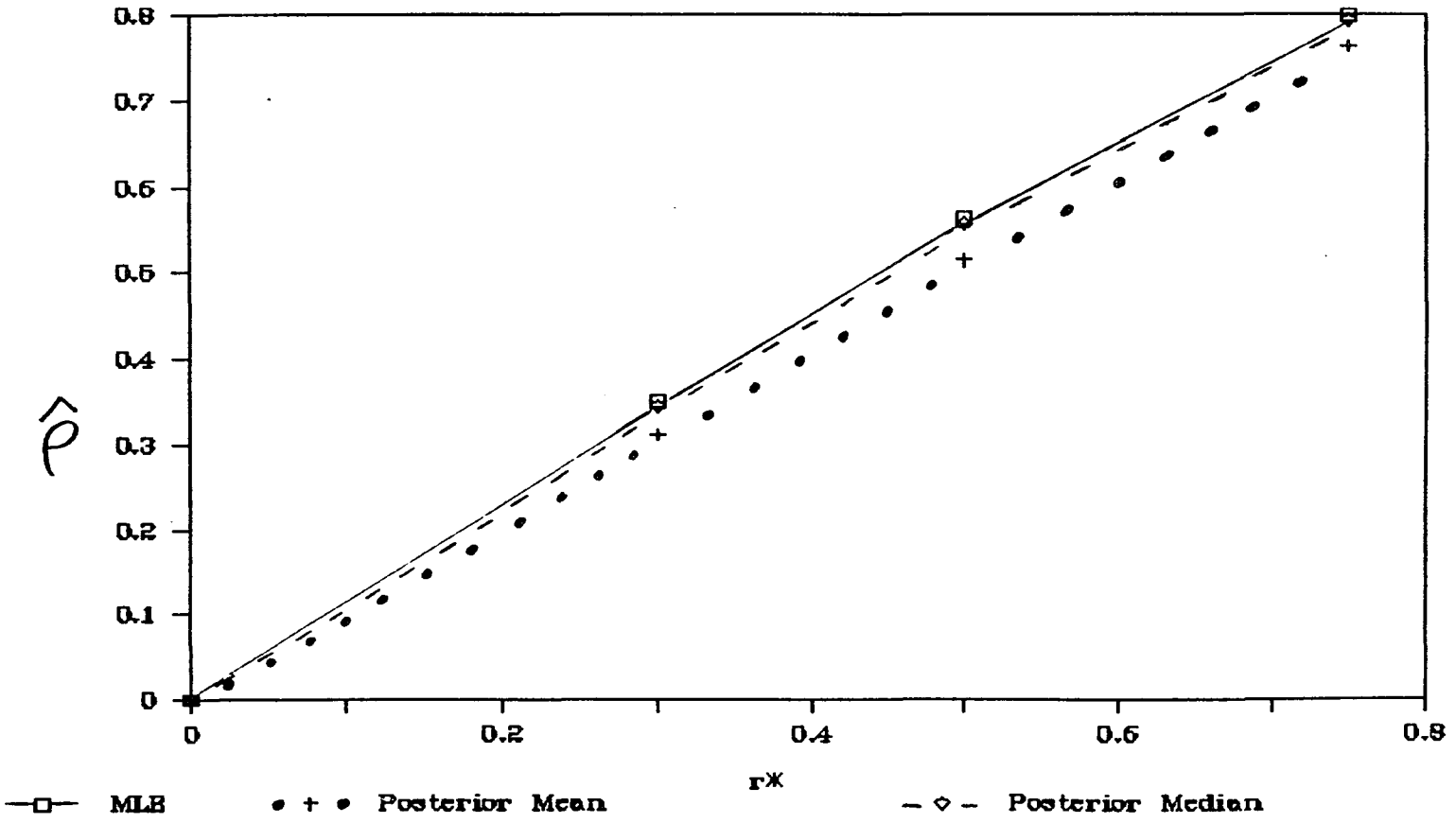


TABLE 1

MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF "n" AND "r*" WHEN $-1 \leq \rho \leq 1$ *

n	r*	MLE	Posterior Mean	Posterior Median
12	.0	.0	.0	.0
12	.3	.349	.263	.335
12	.5	.562	.445	.544
12	.75	.797	.692	.781
20	.0	.0	.0	.0
20	.3	.349	.294	.342
20	.5	.562	.482	.553
20	.75	.797	.741	.790
30	.0	.0	.0	.0
30	.3	.349	.310	.344
30	.5	.562	.514	.556
30	.75	.797	.762	.792

*Note: Figures 4, 5 and 6 are based on the above values.

The results in figures 4, 5 and 6 parallel those in figures 1, 2, and 3 for the interval estimates, in that the Bayesian and maximum likelihood approaches yield similar results when ρ is not restricted. The major conclusions that can be drawn from these results are:

a. The posterior medians and the MLE's are almost equal for all sample sizes.

b. For small sample sizes the posterior mean is smaller than the MLE (see figure 4 where n=12), but as the sample size becomes larger the posterior mean and the MLE become very similar (see figure 6 where n=30).

Figures 7, 8 and 9 show the same information as figures 4, 5 and 6 with the exception that now when computing the posterior means and medians the prior distribution of ρ is given as, $P'(\rho)=(1/1-\rho^2)^{3/2}$ where $0\leq\rho\leq1$. As was shown in figures 1, 2 and 3, the Bayesian and ML estimates diverge when ρ is restricted to be positive. Table 2 presents the values on which figures 7, 8 and 9 are based. The major conclusions that can be drawn from these results (figures 7, 8 and 9) are:

a. The Bayesian estimates tend to be higher than the MLE's. For example, in figure 7 (n=12) for r*=.3 the posterior median is .4815, the posterior mean .475 and the MLE .349.

FIGURE 7
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=12 $0 \leq \rho \leq 1$)

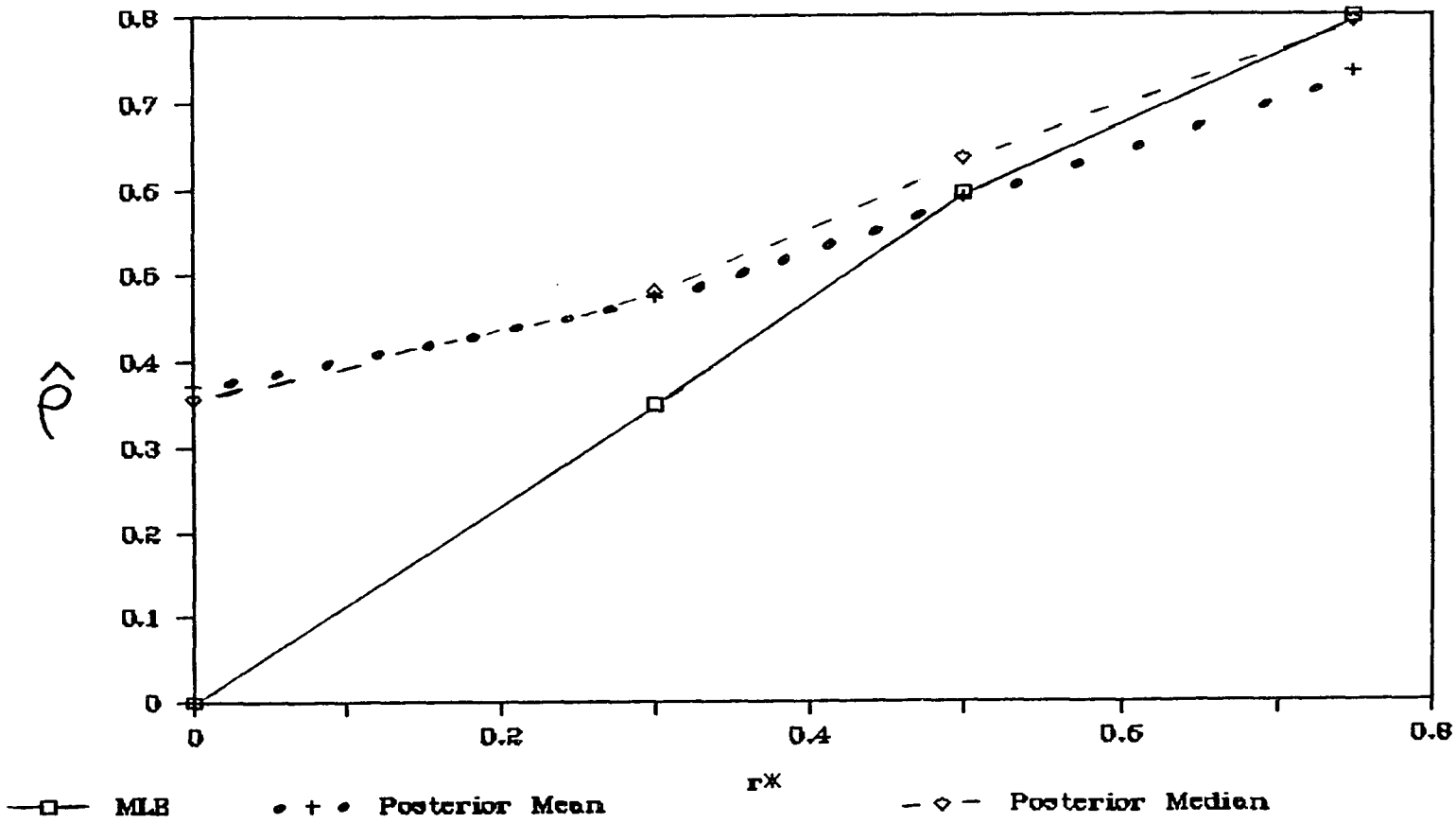


FIGURE 8
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=20 $0 \leq \rho \leq 1$)

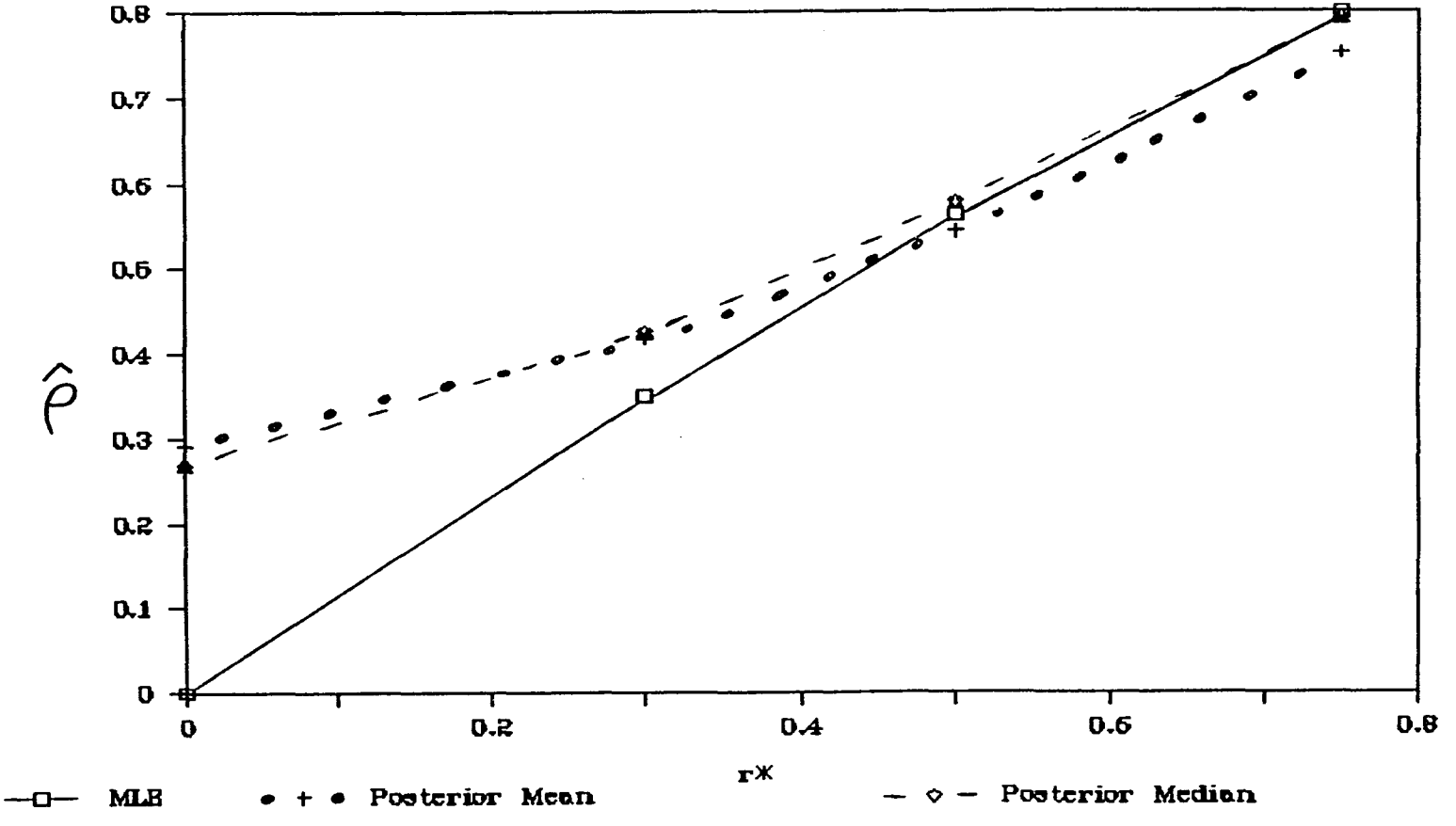


FIGURE 9
MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF r* (n=30 $0 \leq \rho \leq 1$)

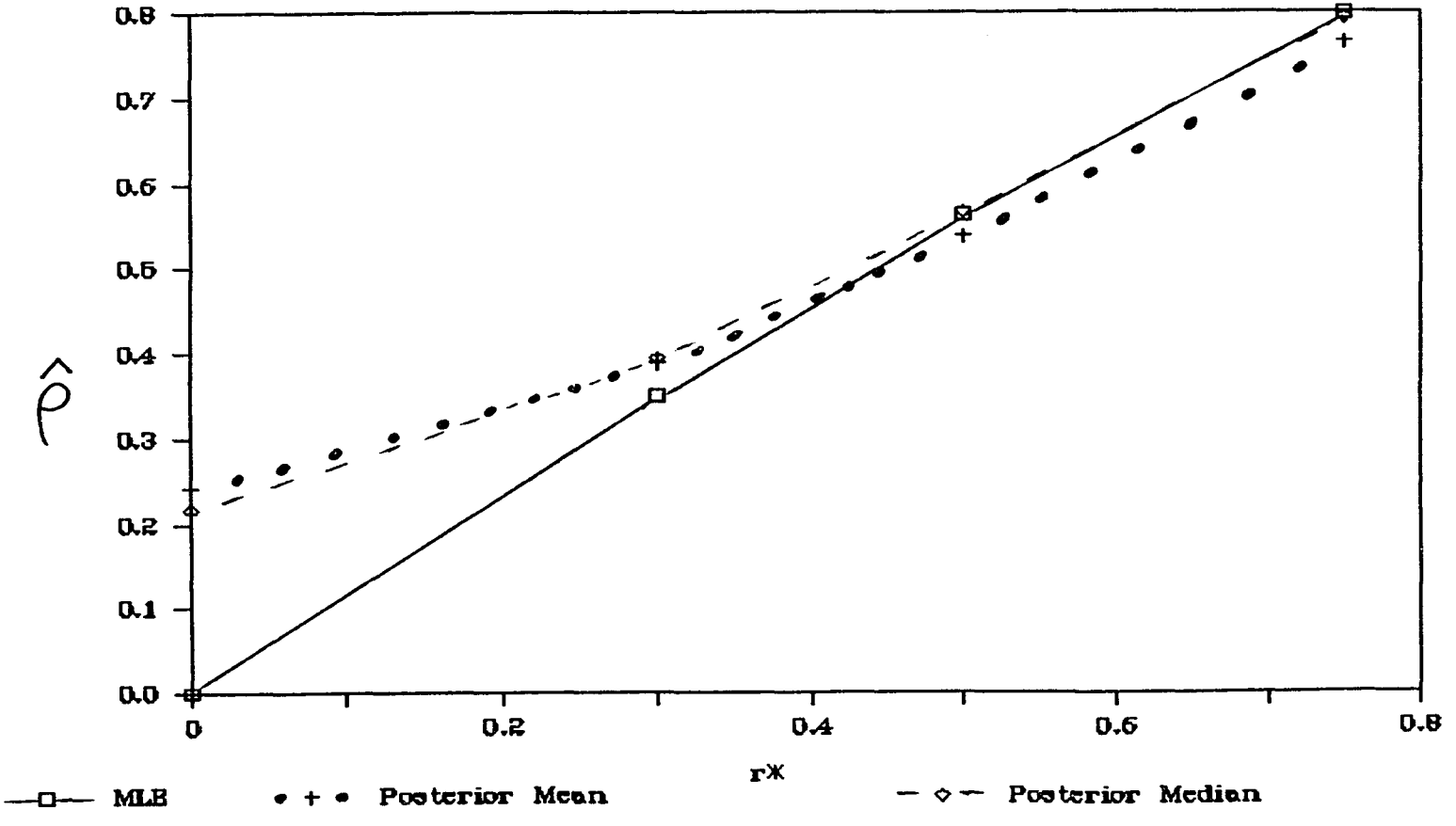


TABLE 2

MLE, POSTERIOR MEAN AND POSTERIOR MEDIAN AS A FUNCTION OF "n" AND "r*" WHEN $0 \leq \rho \leq 1$ *

n	r*	MLE	Posterior Mean	Posterior Median
12	.0	.0	.370	.356
12	.3	.349	.474	.481
12	.5	.562	.589	.635
12	.75	.797	.733	.793
20	.0	.0	.291	.268
20	.3	.349	.417	.425
20	.5	.562	.543	.577
20	.75	.797	.751	.791
30	.0	.0	.241	.216
30	.3	.349	.387	.393
30	.5	.562	.537	.566
30	.75	.797	.764	.793

*Note: Figures 7, 8 and 9 are based on the above values.

b. As the sample size (n) and the r* increase the MLE and the Bayesian estimates become more similar. For example, for the biggest sample size [figure 9 (n=30)], for r*=.3 the MLE is .349, the posterior mean equals .387 and the posterior median .394. In figure 7 (n=12) it can be seen how the estimates become more similar for larger values of r*, i.e. the corresponding lines for the Bayesian estimates and MLE's become closer to each other as r* increases.

C. Factors affecting the Bayesian Estimates

It is also of interest to investigate how the posterior mean and median vary as a function of r*, R_1, R_2 and n. To obtain descriptive information, factorial tables were constructed showing how the mean (median) varies as a function of the three R_1 values, the three R_2 values, the three n values and the four r* values.

More specifically, both the posterior mean and median were analyzed using a four-way ANOVA procedure. The Mean Square values for each factor (n, R_1, R_2, r*) and for the two-way, three-way and four-way interactions were observed. Those effects having relatively large Mean Squares are presented in Table 3. Tables 4 and 5 show the values for the Mean Squares.

TABLE 3

POSTERIOR MEAN AND MEDIAN AS A FUNCTION OF R_1, R_2, n AND r* ($0 \leq \rho \leq 1$)

R_1 (nc/n)	.30	.50	.80	
Posterior				
Mean	.52	.51	.49	
Median	.55	.53	.50	
R_2 (s_x^{*2}/s_x^2)	.50	.75	1.00	
Posterior				
Mean	.56	.51	.46	
Median	.58	.52	.47	
n	12	20	30	
Posterior				
Mean	.54	.50	.48	
Median	.57	.52	.49	
r^*	.00	.30	.50	.75
Posterior				
Mean	.30	.43	.56	.75
Median	.28	.43	.59	.79

TABLE 4

MEAN SQUARES FOR THE FOUR-WAY ANOVA FOR THE POSTERIOR MEAN

	MEAN SQUARE
MAIN EFFECTS	
R1	.008
R2	.083
n	.033
r*	.993
TWO-WAY INTERACTIONS	
R1 by R2	.001
R1 by n	.001
R1 by r*	.012
R2 by n	.000
R2 by r*	.001
n by r*	.011
THREE-WAY INTERACTIONS	
R1 by R2 by n	.001
R1 by R2 by r*	.000
R1 by n by r*	.001
R2 by n by r*	.001
FOUR-WAY INTERACTION	
R1 by R2 by n by r*	.000

Note: $0 \leq \rho \leq 1$

TABLE 5

MEAN SQUARES FOR THE FOUR-WAY ANOVA FOR THE POSTERIOR MEDIAN

	MEAN SQUARE
MAIN EFFECTS	
R1	.017
R2	.107
n	.052
r*	.301
TWO-WAY INTERACTIONS	
R1 by R2	.001
R1 by n	.001
R1 by r*	.009
R2 by n	.001
R2 by r*	.001
n by r*	.008
THREE-WAY INTERACTIONS	
R1 by R2 by n	.001
R1 by R2 by r*	.001
R1 by n by r*	.001
R2 by n by r*	.001
FOUR-WAY INTERACTION	
R1 by R2 by n by r*	.001

Note: $0 \leq \rho \leq 1$

It appeared that only main effects were present for R_1, R_2, n and r*. As expected, the strongest effect on both, the posterior mean and median was for r*. This effect can be seen in Table 3; as r* increases the Bayesian estimates strongly increase.

It is also seen in Table 3 that as R2 (s_x^{*2}/s_x^2) increases, the Bayesian estimates on average decrease. In other words, as the missing data becomes closer to being MCAR (R_1=1), the estimate of ρ on average is lowered. When the data is MAR, the r* is an underestimate of ρ. Thus, the more the data is MAR versus MCAR (smaller R_2), the larger the estimate of ρ should be and vice versa.

Similarly, the estimates of ρ decrease as R_1 increases. The smaller the proportion of missing cases the less r* would tend to underestimate ρ; therefore, the smaller the estimate of ρ should be and vice versa. Similarly, as the sample size (n) increases, the Bayesian estimates on average decrease.

It should be noted that there was a small interaction of R_1 with r*. The effect of R_1 on both estimates was stronger for smaller values of r*, i.e., the rate of decrease in the Bayesian estimates decreased as r* increased. There was also a small interaction of n with r*. As in the previous case, the effect of n was greater for small values of r*.

Chapter V

SUMMARY AND DISCUSSION

This study presents a new approach, a Bayesian approach, to estimate ρ (correlation between two continuous variables) for the cases where one of the variables is either MAR or MCAR. A formula for the posterior distribution of ρ when the data are MAR or MCAR was developed. Next, this Bayesian formula was evaluated as a function of different factors (R_1, R_2, n and r*) and compared to the Maximum Likelihood approach for estimating ρ when data are missing at random or missing completely at random.

Ninety percent HDR's, 90% CI's, posterior means, posterior medians and MLE's of ρ were computed for different values of R_1, R_2, n, and r*. The mean upper and lower bounds of the intervals were obtained by averaging over the R_1 and R_2 values. When the Bayesian approach did not restrict the range of ρ ($-1 \leq \rho \leq 1$), the HDR was very similar to the CI. Neither of them was very precise for small sample sizes. Even for larger sample sizes, the intervals were wide for small r* values.

When ρ was restricted to be positive ($0 \leq \rho \leq 1$), the HDR was narrower than the CI and with higher bounds than those for the CI. However, as the sample size increased, the difference between the CI and the Bayesian interval decreased. Similarly,

as the r* value increased, the difference between the CI and the HDR decreased.

The average MLE's and the average posterior means and medians were then computed by averaging over all values of R_1 and R_2. When ρ was not restricted the Bayesian estimates and the MLE were practically equal except for small sample sizes where the posterior mean was smaller than the MLE. When ρ was restricted to be positive, the Bayesian estimates tended to be higher than the MLE; but as the sample size and the r* values increased, the MLE and Bayesian estimates became more similar.

Both the posterior mean and median were analyzed using a four-way ANOVA procedure in order to obtain descriptive information on how the posterior mean and median varied as a function of r*, R_1, R_2, and n. The Mean Square values for each factor and for the two-way, three-way and four-way interactions were observed.

It appeared that only main effects were present. The strongest effect on both estimates was for r*, such that as r* increased the Bayesian estimates of ρ strongly increased. As R_2 increased the posterior mean and median decreased. Similarly, as R_1 increased the estimates decreased and as n increased the estimates also decreased. These results can be explained as follows: the more the data are MAR versus MCAR (smaller R_2) and the larger the proportion of missing cases

(smaller R_1), the more r* will underestimate ρ; thus, the smaller the R_1 and the smaller the R_2, the larger the estimate of ρ should be.

The Bayesian approach is most advantageous when there is prior information about ρ (i.e., ρ is restricted to a range), because the HDR's are considerably narrower than the CI's. This is specially true for small n's and for small r*'s. For larger n's (n=30) the HDR's and the CI's are more similar. However, when r* is small the Bayesian approach is still more precise than the ML approach even for bigger samples (n=30).

It should be noted that the computation of the Bayesian intervals and estimates is not difficult given the availability of computer routines to perform double numerical integration. In the present paper, this computation was performed using a widely available IMSL double precision subroutine.

The results presented in this paper were all obtained under a single specification for the prior distribution of ρ, i.e., $p'(\rho)=(1/1-\rho^2)^{3/2}$. However, one might consider different choices for the non informative prior, e.g., $p'(\rho)=1$, $a<\rho<b$. By looking at the expression of the posterior distribution for ρ and Ψ it is clear that the effect of the prior distribution chosen in this paper is reduced as n increases. To see this result, consider the numerator of the posterior distribution

of ρ and Ψ $[(1-\rho^2)^{(n-1)/2}\ p'(\rho)]$; if we set $p'(\rho)=(1/1-\rho^2)^{3/2}$, this numerator becomes $(1-\rho^2)^{(n-4)/2}$; if we let $p'(\rho)=1$ the numerator becomes $(1-\rho^2)^{(n-1)/2}$. Thus, when n is large the effect of the choice of the non informative prior will be negligible. In addition, for small values of r*, e.g. $0 \leq r^* \leq .10$, the results will also tend to be insensitive to the choice of the non informative prior distribution. The posterior distribution of ρ [equation 28] can be viewed as the product of $p'(\rho)$ and a second factor. For $p'(\rho)=1$ and r* near zero, the posterior density of ρ will tend to strongly be skewed to the right with near zero densities for high values of ρ. Even though the prior $p'(\rho)=(1/1-\rho^2)^{3/2}$ assigns high density values to high values for ρ, this prior density will be dominated by the second factor in equation 28.

Nevertheless, there will most likely be cases where the specific non informative prior chosen for ρ will have an effect on the Bayesian estimates and intervals. For example, consider the effect of choosing as the prior $p'(\rho)=1$ versus $p'(\rho)=(1/1-\rho^2)^{3/2}$ when n is small. However, if we decrease the upper range for ρ (e.g., $0 \leq \rho \leq .8$) the Bayesian estimates and intervals will tend to be similar for both priors. This is so because the difference between these two distributions [$p'(\rho)=1$ and $p'(\rho)=(1/1-\rho^2)^{3/2}$] occurs mostly for large values of ρ.

One additional point should be noted concerning the generalizability of our results. The Fisher Z transform we have employed is strictly speaking only applicable for the complete data case. A different transform is needed for the missing data case. It can be shown that by using the complete case transform, we have underestimated the width of the Maximum Likelihood intervals. Thus, the Bayesian approach is even more advantageous than it appears in the present paper.

There are some unanswered questions which still remain to be considered in future studies. Firstly, It would be of interest to compared the Bayesian and the MLE approaches in those situations where the non informative prior distribution used for ρ place both a lower and an upper bound on ρ (e.g. $.1<\rho<.7$). The results obtained in the present paper suggest that in these cases the Bayesian approach would be even more advantageous. That is, the more restricted the non informative prior distribution, the more precise the estimation is likely to be.

Secondly, different methods can be considered for comparing the Bayesian and MLE approaches. In the present study the data were fixed at specific values and ρ was treated as a random variable. However, given that a proper prior distribution is assigned to ρ, both ρ and the data could be treated as random variables. For example, one could sample a value of ρ from the proper prior distribution of ρ, draw a set

of incomplete samples given ρ, and finally, compute CI's, HDR's, MLE's, posterior means and posterior medians. This procedure would be repeated a specified number of times for different values of ρ that would be sampled from the proper prior distribution assigned to ρ.

Thirdly, a useful extension of the present study is to consider the case where data are missing on both variables X and Y; where X is MCAR or MAR with respect to Y and Y is MCAR or MAR with respect to X. In this case the data set for n subjects could be partitioned into three data sets consisting of n_1, n_2 and n_3 subjects respectively ($n = n_1+n_2+n_3$):

Subj.	X	Y
1	x_1	y_1
2	x_2	y_2
.	.	.
.	.	.
.	x_{n1}	y_{n1}
.	x_{n1+1}	?
.	x_{n1+2}	.
.	.	.
.	x_{n1+n2}	?
.	?	$y_{n1+n2+1}$
.	.	.
.	.	.
n	?	y_n

It is seen that the data show three patterns: complete data on X and Y, data only on X, and data only on Y. The development of a formula for the posterior distribution of ρ in this case would involve a sequential procedure. The first step would consist of obtaining the posterior distribution of $\theta^*=\mu_x$, γ_x, β_0, β_1, γ_y (where $\gamma_x=1/\sigma_{xx}$ and $\gamma_y=1/\sigma_{yy|x}$) using the complete cases and the cases for which X is observed but Y is

missing. In a second step one would incorporate the cases for which the Y variable is observed but for which X is missing. This step would involve the reparametarization of $\theta*$ into $\theta^{+}=\mu_y,\ \gamma_y^{+},\ \gamma_x,\ \beta_0,\ \beta_1$; where $\gamma_y^{+}=1/\sigma_{yy}$. Given the joint posterior of θ^{+}, one could obtain the marginal posterior for ρ.

Future research should also address a more general case, the case where one wants to estimate the multiple correlation between Y and several predictors, given a data set containing missing data on possibly all variables. In this situation, it is probably necessary to use imputation techniques suggested by Tanner and Wong (1987) in order to obtain the Bayesian estimates and intervals.

References

Alexander, R.A., Alliger, G.M., & Hanges, P.J. (1984). Correcting for range restriction when the population variance is unknown. Applied Psychological Measurement, 8, 431-437.

Alexander, R.A., Barrett, G.V., Alliger, G.M., & Carson, K.P. (1986). Towards a general model of non-random sampling and the impact on population correlation: generalizations of Berkson's fallacy and restriction of range. British Journal of Mathematical and Statistical Psychology, 39, 90-105.

Alexander, R.A., Carson, K.P., Alliger, G.M., & Barrett, G.V. (1984). Correction for restriction of range when both X and Y are truncated. Applied Psychological Measurement, 8, 231-241.

Alexander, R.A., Carson, K.P., Alliger, G.M., & Carr, L. (1987). Correcting doubly truncated correlations: an improved approximation for correcting the bivariate normal correlation when truncation has occurred on both variables. Educational and Psychological Measurement, 47, 309-315.

Alexander, R.A., Hanges, P.J., & Alliger, G.M. (1985). Correcting for restriction of range in both X and Y when the unrestricted variances are unknown. Applied Psychological Measurement, 9, 317-323.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. Journal of Applied Psychology, 68, 584-589.

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. Applied Psychological Measurement, 4, 385-398.

Cohen, A.C., Jr. (1955). Restriction and selection in samples from bivariate normal distributions. Journal of the American Statistical Association, 50, 884-893.

Cohen, A.C., Jr. (1957). Restriction and selection in multinormal distributions. Annals of Mathematical Statistics, 28, 731-741.

Cohen, A.C. (1959). Simplified estimators for the normal

distribution when samples are singly censored or truncated. Technometrics, 1, 217-237.

Dobson, P. (1988). The correction of correlation coefficients for restriction of range when restriction results from the truncation of a normally distributed variable. British Journal of Mathematical and Statistical Psychology, 41, 227-234.

Gross, A.L. (1982). Relaxing the assumptions underlying corrections for restriction in range. Educational and Psychological Measurement, 42, 795-801.

Gross, A.L., & Fleischman, E.L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. Applied Psychological Measurement, 7 (2), 227-237.

Gross, A.L., & Fleischman, E.L. (1987). The correction for restriction of range and nonlinear regressions: a analytic study. Applied Psychological Measurement, 11, 211-217.

Gross, A.L., & McGanney, M.L. (1987). The restriction of range problem and nonignorable selection processes. Journal of Applied Psychology, 72, 604-610.

Gross, A.L., & Kagen, E. (1983). Not correcting for restriction of range can be advantageous. Educational and Psychological Measurement, 43, 389-396.

Gross, A.L., & Perry, P. (1983). Validating a selection test, a predictive probability approach. Psychometrika, 48, 1, 113-127.

Greener, J.M., & Osburn, H.G. (1980). Accuracy of corrections for restriction in range due to heteroscedastic and non-linear distributions. Educational and Psychological Measurement, 40, 337-346.

Holmes, D.J. (1990). The robustness of the usual correction for restriction in range due to explicit selection. Psychometrika, 55, 19-32.

Lee, P.M. (1989). Bayesian Statistics: An introduction. New York: Oxford University Press.

Lee, R., Miller, K.J., & Graham, W.K. (1982). Corrections for restriction of range and attenuation in criterion-

related validation studies. Journal of Applied Psychology, 67, 637-639.

Lindley, D.V. (1965). Introduction to probability and statistics from a Bayesian viewpoint. Part 2: Inference. Cambridge at the University Press.

Linn, R.L. (1968). Range restriction problems in the use of self-selected groups for test validation. Psychological Bulletin, 69, 69-73.

Linn, R.L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. Journal of Educational Measurement, 20, 1-15.

Maddala, G.S. (1983). Limited, dependent and qualitative variables in econometrics. New York: University Press.

Mendoza, J.L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. Journal of Educational Statistics, 12, 282-293.

Roe, R.A. (1979). The correction for restriction of range and the difference between intended and actual selection. Educational and Psychological Measurement, 39, 551-559.

Roe, R.A., & Elshout, J.J. (1972). Some new formulas for the correction for restriction of range. Nederlands Tijdschrift voor de Psychologie, 27, 134-139.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. (1976). Statistical power in criterion-related validation studies. Journal of Applied Psychology, 61, 473-485.

Tanner, M.A., & Hung Wong, W. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82, 398(number), 528-540.

Thorndike, R.L. (1947). Research problems and techniques (Report No.3) Washington DC: AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office.

Wells, D.G., & Fruchter, B. (1970). Correcting the correlation coefficient for explicit restriction in both variables. Educational and Psychological

Measurement, 30, 925-934.