



**UNIVERSITY OF
GÖTEBORG**
Department of Statistics

RESEARCH REPORT 1991:1

ISSN 0349-8034

ON SOME PREDICTION METHODS
FOR CATEGORICAL DATA

by

Jonny Olofsson

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

ACKNOWLEDGEMENT

The author is grateful to Associate Professor Marianne Frisé, who suggested the topic for the paper and provided considerable encouragement and valuable comments during the course of the work.

He also wishes to thank Mr Christer Andersson for carefully reading the manuscript and Mr Staffan Geissler who printed the manuscript.

ABSTRACT

Good prediction methods are important in many fields where qualitative variables are involved. The criterion of a good prediction method, used in this paper, is the average mean squared error. This criterion is used to compare and derive prediction methods, when the variable of interest is binary. The methods considered here are based on the maximum-likelihood estimators of the expectation of the binary variable, for which we want to make a prediction. Derivations and simulations are made for the case where we have one qualitative background variable. It is for example demonstrated that, when the ordinary chi-squared test is used for choosing between two prediction methods, it should not be adopted on a conventional low level of significance (e.g. 5%).

CONTENTS

1 Introduction	1
2 Different kinds of predictions and measures of prediction error	5
2.1 Event predictions and actuarial predictions	5
2.2 Measures of prediction error	7
2.3 Model selection procedures	14
3 Models and notations	18
4 Derivation and comparisons of prediction rules	25
4.1 The average mean squared error for the predictors based on the unrestricted and the restricted model	25
4.2 The suggested predictor P_b^{\sim} .	32
4.3 Comparison of P_b^{\sim} and predictors based on Pearson's chi-squared test	35
4.4 Comparison of P_b^{\sim} and predictors based on the Akaike-criterion	41

4.5 Effects of different sampling proportions in the old and the new sample	44
4.6 The multinomial case	49
4.7 Figures	52
5 Concluding remarks	65
Appendix	66
References	

1. INTRODUCTION

In this paper we consider model selection, when faced with a binary dependent variable, Y , and a number of qualitative background variables. In an application Y could, for example, correspond to the presence or absence of a certain disease, and the background variables could be, e.g. exposed/not exposed, sex, living area. The data can be organized in a multi-dimensional contingency table, where each cell contains the number of observations for a certain combination of variable values.

Even for a moderate number of background variables, there are a large number of cells in the table, and therefore there is often a wish to reduce, if possible, the dimensionality of the table. Traditionally this reduction has been accomplished by fitting various log-linear models to the data and removing parameters from the model that have proved non-significant according to some kind of statistical test, thus obtaining a non-saturated model, i.e. a model containing fewer parameters than cells in the table. Section 2.3 gives a brief discussion on such procedures. By estimating the parameters of a particular model, we can also obtain estimates of the conditional probability of Y given the values of the background variables.

It is not self-evident that the traditional strategy, involving testing of hypotheses, is the best for all purposes. The

purpose of the analysis can, for example, be to obtain information about the causal structure or to obtain predictors that will minimize some measure of error. We will in this paper concentrate on the latter aim, restricting the analysis to the class of predictors, where the maximum-likelihood (ML) estimates replace the parameters of a 'good' model. The problem is thus to find 'good' models for prediction purposes. This approach has similarities with the traditional ones and might, besides giving good predictors, also give some insight in to the structure of the data. The principal theoretical differences between the statistics suitable for traditional testing of hypotheses and those suitable for prediction purposes, are also of interest.

The problem of choosing models for prediction purposes has been extensively studied in the area of multiple regression analysis. A criterion for model selection frequently adopted is the mean squared error of prediction (MSEP). For observation of a future value of a binary dependent variable Y , MSEP is defined as $E(Y-p')^2$, where p' is a predictor of Y . As the MSEP takes account of both bias and sampling variability, we have the result that the saturated model is not necessarily the best. A bias can very well be offset by a reduced sampling variability owing to the inclusion of fewer estimated parameters. The MSEP depends on parameters which usually have to be estimated.

It must be noted here, that we do not argue that the selected model is the "true" one. That is, in selecting a non-saturated

model, we have not proved that the remaining effects are equal to zero. All we can say is that the selected model has the best prediction ability as judged by the criterion used in the study (with the reservation that the criterion is estimated on the basis of data).

Chapter 2 presents a brief discussion on two different kinds of prediction methods for binary data. Different measures of prediction error are also considered. The chapter ends with a short review of some procedures for selection of models.

Chapter 3 introduces notations and a measure of prediction error, for the case where we want to make predictions about a binary variable and where we have observations on a discrete background variable, Z . When making a prediction for a particular level of Z , we distinguish between two predictors. The first is the usual maximum-likelihood estimator of the probability of success and the second is the maximum-likelihood estimator, which is obtained under the restriction that all success probabilities are equal (i.e. homogeneity). In the following these predictors will be referred to as the unrestricted and restricted predictor, respectively.

In Chapter 4, we examine a measure of prediction error, the average mean squared error, AMSE, for both prediction rules. A criterion based on AMSE for choosing between the two prediction rules is developed. The AMSE for this combined prediction rule is compared with the AMSE for the rules P' and P^* . We also

study the AMSE of the prediction rule that is obtained by letting a chi-squared test of homogeneity make the selection between P' and P^* . The AMSE-criterion is also compared with the so called Akaike-criterion.

2. DIFFERENT KINDS OF PREDICTIONS AND MEASURES OF PREDICTION ERROR

2.1 Event predictions and actuarial predictions.

When making predictions of a variable, we often use known values of other variables, in some way related to the unknown variable. The variable for which we want a prediction is called the dependent variable, while the other (background-) variables are termed independent. In this section we discuss some distinctions among alternative kinds of predictions, when both the dependent and the independent variables are categorical. As the prediction ability of a specific model can be used as a criteria for model selection, we will also look at some measures of prediction error.

Hildebrand et al. (1977) make a distinction between an event prediction and an actuarial prediction. An event prediction is a proposition that predicts each case's state on the dependent variable, while an actuarial prediction is a proposition which specifies, for each case, the probabilities of the dependent variable. As an example of an event prediction rule the authors take the case of a binary dependent variable and two independent variables:

"If the legislator is liberal from an urban district then predict that that person will vote in favour of the bill"

An example of an actuarial prediction is:

"The chance of rain during each day in July is $1/3$ "

In the first case an investigator could determine, for any "liberal from an urban district", whether the prediction was correct once the vote has been cast. In the second case the investigator could evaluate the extent to which the observed proportion matched the predicted. Usually, both kinds of predictions are based on past data. Having noted a proportion of $1/3$ rainy days in July over a number of years, we could make the prediction "no rain" for a certain day in July, because this alternative is the most likely.

In many cases it seems reasonable to associate an actuarial prediction with an individual future observation. Instead, for example, making an event prediction and classifying a patient as either healthy or sick, we come up with a proportion reflecting the risk of having the disease. Increasing values of the prediction could perhaps correspond to actions reaching from surveillance, via drug treatment, up to surgical treatment.

2.2 Measures of prediction error.

The topic covered in this paper, is a special case of a more general problem where one wants to choose a model suitable for prediction purposes. See Linhart and Zucchini (1986) for examples of application in different fields. In the literature the measure of prediction error is often termed *error rate* and one distinguishes the *optimum error rate*, which is the error rate that can be obtained if the parameters of the statistical model are known and the optimal predictor is used. Secondly, the *actual error rate*, is the error rate obtained by averaging over the the distribution of future observations. Thirdly, the *apparent error rate* is defined to be the average error rate when the predictor is applied to the available observations retrospectively. A trivial example should illuminate these concepts.

Suppose we observe a sequence of independent random variables Y_1, \dots, Y_n , with common mean μ and variance σ^2 . Suppose that we want to use these observations to make a prediction of a future observation Y_{new} , thought to have the same distribution. Now, suppose we adopt the mean squared error to a particular predictor Y' , for example the mean of the observations Y_1, \dots, Y_n . The actual error rate then becomes

$$\text{MSE}_{\text{act}} = E(Y_{\text{new}} - Y')^2 = (\mu - Y')^2 + \sigma^2$$

If the parameter μ were known, we could use it as a predictor and we would obtain the optimum error rate

$$\text{MSE}_{\text{opt}} = \sigma^2$$

For the apparent error rate we let the predictor Y' predict the observations retrospectively and average the squared errors

$$\text{MSE}_{\text{app}} = \Sigma (Y_i - Y')^2/n$$

Much research has been devoted to estimating the expectation of actual error rate (the apparent error rate is generally biased downward). Van Houwelingen and Le Cessie (1989) gives a review of different ways for estimation, including cross-validation. Efron (1986) provides several estimates for the bias of the apparent error rate. The theory applies to general exponential family linear models and general measures of prediction error.

In a setting identical to the one in this paper, where there are several groups of observations on a binary variable, Efron (1978), constructs one-way ANOVA tables, by introducing a wide class of measures of binary variation, including the squared error. A coefficient of determination can thus be defined for each measure in the class, reflecting the proportional decrease in residual variation when going from a crude explanation of the probabilities of success in the groups, to a more detailed.

In the case of event predictions, we are either right or wrong, so one appropriate measure of prediction error is the apparent error rate:

The number of false predictions divided
by the total number of predictions

Consider the following example, where we have some past data of how urban legislators voted in a similar election.

	Liberal	Conservative
In favour	10	50
Against	30	10

Suppose we adopt the strategy of predicting "against" for liberals and "in favour" for conservatives. The apparent error rate in this case equals 0.20 and can be interpreted as follows. Suppose we knew only the state of the independent variable for all 100 individuals and used the proposed strategy to predict the state of the dependent variable of a randomly selected individual. The probability of making a false prediction would then be 0.20. This rate could be used for comparing other prediction rules, e.g. predicting "in favour" for both liberals and conservatives.

Since the apparent error rate was obtained by letting the sample predict itself, we might suspect that it is too optimistic for future data. That this is the case is shown in Efron (1986).

Turning to actuarial predictions where the predictor is continuous on the interval $(0,1)$, we are more flexible when choosing a measure of prediction error. In the setting of the two-way classification of above, we define the squared prediction error for a future observation, i , belonging to state j on the independent variable

$$(p_j' - Y_{ij})^2$$

where p_j' is an estimate of the probability $P(Y=1|Z=j)$. The actual error rate in this case is

$$\begin{aligned} E(p_j' - Y_{ij})^2 &= p_j'^2 - 2 \cdot p_j' \cdot p_j + p_j = \\ &= (p_j' - p_j)^2 + p_j \cdot (1 - p_j) \end{aligned}$$

When the aim is to compare the performance of different predictors, we could of course drop the constant term

$p_j \cdot (1-p_j)$. In chapter 4, we will study the expectation of a weighted average of this error rate over the values of j for two different predictors.

Another measure is the expectation of the Kullback-Leibler distance. For a single Bernoulli variable Y with expectation p this is defined as:

$$D_{\text{act}} = E(-Y \cdot \log(p') - (1-Y) \cdot \log(1-p')) = \\ -p \cdot \log(p') - (1-p) \cdot \log(1-p')$$

where p' is a predictor of Y and the expectation is taken over Y , holding p' constant. It is equal to the expectation of the log-likelihood over Y , holding p' constant. We see that for $Y=1$ the Kullback-Leibler distance is equal to $-\log(p')$, a decreasing function of p' and for $Y=0$ it is equal to $-\log(1-p')$, an increasing function of p' . The apparent error is

$$D_{\text{app}} = -p' \cdot \log(p') - (1-p') \cdot \log(1-p')$$

If we in the two-way classification assume that we observe one binomial variable X_j for each level of the independent variable, with parameters (n_j, p_j) $j=1,2,\dots,k$, the Kullback-Leibler discrepancy can be written as

$$\begin{aligned}
 D_{\text{act}} &= E(- \sum X_j \cdot \log(p_j') - \sum (n_j - X_j) \cdot \log(1 - p_j')) = \\
 &= - \sum p_j \cdot \log(p_j') - \sum (n_j - X_j) \cdot \log(1 - p_j')
 \end{aligned}$$

where p_j' $j=1,2$, are predictors of Y_{ij} . The apparent error becomes

$$D_{\text{app}} = - \sum p_j' \cdot \log(p_j') - \sum (n_j - p_j') \cdot \log(1 - p_j')$$

Now, the difference $D_{\text{act}} - D_{\text{app}}$ can be written as

$$\sum (p_j' - p_j) \cdot \log(p_j' / (1 - p_j'))$$

Approximating $\log(p_j' / (1 - p_j'))$ with the first two terms in it's Taylor expansion, the expectation of $D_{\text{act}} - D_{\text{app}}$ is

$$\sum E((p_j' - p_j) \cdot \log(p_j' / (1 - p_j)) + (p_j' - p_j)^2 / p_j(1 - p_j))$$

If we for example let p_j' be the ordinary ML-estimator of p_j , this expectation is equal to $2/n$. If we have k different binomial populations the expectation would be k/n . Adjusting D_{app} with the bias approximation we get

$$\tilde{D} = D_{\text{app}} + 2/n$$

This in turn is equal to

$$- l + 2/n$$

where l is the maximized log-likelihood.

In fact, this is a special case of the generalized information criterion for model selection, which states that one should choose the model for which

$$l_i - 0.5 \cdot \alpha \cdot q_i$$

is maximum. l_i is here the log-likelihood for the i :th model, maximized over q_i parameters. In our case α is equal to 2 and this corresponds to the Akaike information criterion. For a discussion of this and the generalized information criterion see Atkinson (1980) and section 4.4.

2.3 Model selection procedures.

When we have decided on a particular measure to compare models, there are different ways to search for the "best" model. For multidimensional contingency tables, one often considers the class of log-linear models, where it is assumed that the logarithms of the cell probabilities depend additively on a number of so called effects. For a three-dimensional table a log-linear model can be written

$$\begin{aligned} \log(p_{ijk}) = & \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + \\ & + (\alpha\beta\delta)_{ijk} \end{aligned}$$

where $p_{ijk} = P(Y=i, Z_1=j, Z_2=k)$ $i=0,1$ $j=1,2,\dots,J$ $k=1,2,\dots,K$

This is the saturated model, e.i. it contains as many effects as there are cells in the table. Unsaturated models are obtained by removing effects. It is a common practice to restrict attention to a family of submodels, called hierarchical models. The hierarchical principle means that if an effect is set equal to zero, then all its higher-order relatives are also set equal to zero. In the three-dimensional case, if for example the second-order interaction $(\alpha\beta)_{ij}$ is zero, the hierarchical principle means that the third-order interaction $(\alpha\beta\delta)_{ijk}$ is also zero. As we are more interested in modelling p_{jk} , the probability of $Y=1$ given the values of the

independent variables, we note that the logit of this probability can be expressed as the difference between two log-linear models

$$\begin{aligned} \log(p_{1jk}) - \log(p_{0jk}) &= \log(p_{jk}/(1-p_{jk})) = \\ &= \alpha_1 - \alpha_0 + (\alpha\beta)_{1j} - (\alpha\beta)_{0j} + (\alpha\delta)_{1k} - (\alpha\delta)_{0k} \\ &\quad + (\alpha\beta\delta)_{1jk} - (\alpha\beta\delta)_{0jk} \end{aligned}$$

As the number of possible models increases rapidly with the number of variables in the table, many model selection procedures have been developed. These procedures end up with one or several models hoped to be adequate in some way.

Most strategies for model selection begins by the fitting of a starting model. Adopting a rule for stepping from one model to another, one searches over a subset of the possible models. The process stops when some termination criterion is fulfilled.

In most selection procedures the stepping rule and the termination procedure, depend on a goodness-of-fit test. Two commonly used test statistics are the Pearson chi-squared statistic and the log-likelihood ratio statistic.

The selection procedures can be divided into three types, which all have a counterpart in multiple regression analysis. Starting from a simple model (often consisting of the main effects only), one conducts *forward stepping* by successively including effects. In *backward stepping* one starts with a complex model (often the saturated model), and successively removes effects. Greater flexibility is obtained if we allow effects previously added to the model to be removed in a later state or allow an effect removed in an earlier state to be included again. Virtually all procedures end when the tests employed for the addition of a term are nonsignificant or the tests for the removal of a term are significant. Benedetti and Brown (1978) summarize several of these procedures and illustrates their performance by an example.

If the aim is to make hypothesis tests of effects, there is a problem of controlling the overall significance level. Aitkin (1979) has developed a simultaneous test procedure for fitting models, which is based on a backward stepping procedure.

Fowles, Freeney and Landwehr (1988) construct a scatterplot for the d.f. (degrees of freedom) versus the value of the log-likelihood statistic for all possible models. Points that fall near the line (d.f., d.f.), fit the data well since the expectation of the test statistic equals the number of degrees of freedom if the model is correct and the sample size is large. Points to the right of the plot (large d.f.) represent simple

models. Suggestions are made to select a subset of models with high d.f.:s near the line (d.f.,d.f.) for further inspection. This is an analogy to the Mallows's C_p -plot for multiple regression.

3. MODELS AND NOTATIONS.

From now on we will be concerned with some simple special cases of choosing a model for a cross-classification when the objective is to make predictions of a binary variable, Y . We assume that we have observations on Y and an attribute Z , which is purely nominal. The data can be presented in a contingency table

		Z			
		1	2	3.....k	
Y	0	x_{01}	x_{02}	$x_{03}.....x_{0k}$	$x_{0.}$
	1	x_{11}	x_{12}	$x_{13}.....x_{1k}$	$x_{1.}$
		$x_{.1}$	$x_{.2}$	$x_{.3}.....x_{.k}$	$x_{..} = m$
		$=m_1$	$=m_2$	$=m_3.....=m_k$	

where x_{ij} is the number of observations in cell (i,j) , and summation over an index is indicated by a dot. The aim is to use these data when making predictions for future values of Y .

The corresponding notations for the probabilities will be

		Z			
		1	2	3.....k	
Y	0	p01	p02	p03.....p0k	p0.
	1	p11	p12	p13.....p1k	p1.
		p.1	p.2	p.3.....p.k	p..= 1

Defining

$$p_j = E(Y | Z=j) = \frac{p_{1j}}{p_{0j}+p_{1j}} \quad j = 1, 2, \dots, k$$

we want to obtain an estimate of each p_j and use it as a predictor of future observations on the binary variable Y. Thus we are dealing with actuarial prediction. This is also the simplest example of variable selection, where we have only one independent variable.

For the two-way classification above we can formulate a saturated log-linear model as

$$l_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where $l_{ij} = \log(p_{ij})$ and $\sum \alpha_i = \sum \beta_j = \sum \sum (\alpha\beta)_{ij} = 0$. For situations where one variable can be interpreted as response and the other as explanatory, log-linear models that condition on the margins of the explanatory variable, that is, logistic regression models are of interest.

$$\begin{aligned} l_{1j} - l_{0j} &= \log(p_{1j}/p_{0j}) = \\ &= \alpha_1 - \alpha_0 + (\alpha\beta)_{1j} - (\alpha\beta)_{0j} \\ &= \mu' + \beta_j' \end{aligned}$$

We assume that the x_{ij} :s are distributed as independent binomial variates, which is equivalent to assuming that the marginals $x_{.j}$:s are fixed. In section 4.6 we will look at the case where only the total sample size is fixed, e.i. the x_{ij} :s are multinomially distributed. The relation $\beta_1' = \beta_2' = \dots = \beta_k' = 0$ corresponds to equality of the probabilities p_j , e.i. homogeneity. Thus we have two possible logistic models, one including only the constant term μ' , the other also including the effect β_j' .

Restricted model: $\log(p_j/(1-p_j)) = \mu'$

Unrestricted model: $\log(p_j/(1-p_j)) = \mu' + \beta_j'$

A selection procedure in this case, thus amounts to choosing between these two model. A common approach is to adopt some test of the hypothesis $H_0: \beta_j' = 0$ and choose the larger model if this test is rejected. As this paper deals with prediction of a binary variable, we want to obtain estimates of p_j :s under both models. Generally, to derive M.L.-estimates of the effects in a logistic model, we need iterative methods. Replacing μ' and β_j' with their estimated values, we can get the M.L.-estimates for the p_j :s. For the restricted model we obtain a single estimate for all p_j :s, while for the unrestricted model they generally differ. In this case, where we have just one independent variable and thus no interaction effects, we can compute the M.L.- estimates directly, without fitting a logistic model.

Now, assuming the restricted model is correct (e.i. homogeneity), the M.L.-estimator of the p_j :s is the sum of the number of successes for each level of Z , divided by the total number of observations. We make the notation

$$p^* = X_{1.}/m$$

If the unrestricted model is correct , we obtain the M.L.-estimators

$$p_j' = X_{1j}/x_{.j} = X_{1j}/m_j$$

i.e. the number of successes divided by the number of observations for $Z=j$. p^* and p_j' will be referred to as the unrestricted and restricted estimator/predictor, respectively. In the following we let p_j' and p^* denote both the stochastic variable and a particular realization.

As a measure of prediction error, we will adopt squared error, that is

$$(p_j' - Y_{ij})^2 \quad \text{and} \quad (p_j^* - Y_{ij})^2$$

where Y_{ij} is a new observation independent of the X_{ij} :s but identically distributed. Following the terminology of chapter 2, the actual error rates are obtained by taking expectation over Y_{ij} , holding p_j' and p^* constant.

$$E(p_j' - Y_{ij})^2 = (p_j' - p_j)^2 + p_j \cdot (1-p_j)$$

$$E(p^* - Y_{ij})^2 = (p^* - p_j)^2 + p_j \cdot (1-p_j)$$

Comparing the performance of the two predictors, we drop the common term $p_j \cdot (1-p_j)$ and take expectation also over p_j' and p_j^* respectively. We make the notation

$$\text{MSE}(p_j') = E(p_j' - p_j)^2 \quad \text{and}$$

$$\text{MSE}(p_j^*) = E(p^* - p_j)^2$$

where MSE stands for 'mean squared error'. Of course, both MSE:s depend on unknown parameters and have to be estimated.

Now, if we make a new observation for which $Z=j$, we would prefer the predictor p_j' if

$$\text{MSE}(p_j^*) - \text{MSE}(p_j') > 0$$

Turning to the case where we want to predict a whole sample of new observations, with sampling proportions f_1, f_2, \dots, f_k , we assume that the aim is to select a vector of predictors that is on the average good for the whole new sample. We will consider two such vectors, namely

$$P' = (p_1', p_2', \dots, p_k')$$

$$P^* = (p_1^*, p_2^*, \dots, p_k^*)$$

both of dimension k . Thus for the prediction rule P' ; if $Z=j$ for an observation, use p_j' for prediction. And for the rule P^* ; if $Z=j$ use p_j^* for prediction. We define the average mean squared error for the two prediction rules as

$$\text{AMSE}(P') = \Sigma f_j \cdot \text{MSE}(p_j')$$

$$\text{AMSE}(P^*) = \Sigma f_j \cdot \text{MSE}(p_j^*)$$

Thus, we would prefer the prediction rule P' if

$$\text{AMSE}(P^*) - \text{AMSE}(P') > 0$$

In section 4.1, we will see that this criterion implies that we should use P^* also for some departures from homogeneity. As the criterion depends on unknown parameters, we will also study the effects of estimating it from the data.

4. Derivation and comparisons of prediction rules.

4.1 The average mean squared errors for the predictors based on the unrestricted and the restricted model.

In this section we study the average mean square error criteria for choosing between the two predictors P' and P^* . We should use the predictor P' for a sample of new observations if

$$AMSE(P') - AMSE(P^*)$$

is larger than zero and otherwise using the predictor P^* . Of course, this criterion depends on unknown parameters, which in practice have to be estimated. Before we turn to this, we will examine how the AMSE:s depend on the true parameters. We will also study the shape of the region in the parameter space where P' is preferred to P^* .

Now, recall that we have available observations x_{1j} with corresponding sample sizes m_j , $j=1,2,\dots,k$. We want to use the old data set to make a prediction rule for new observations from a population with the same probabilities p_j , but with possibly different sampling proportions. The risk of the unrestricted predictor P' can be written as

$$\text{AMSE}(P') = \sum f_j \cdot E(p_j' - p_j)^2 = \sum f_j \cdot p_j \cdot (1-p_j) / m_j$$

Until section 4.5 we assume that the sampling proportions in new sample are equal to the sampling proportions in the old sample, i.e. $f_j = m_j / m$ for $j=1,2,\dots,k$. $\text{AMSE}(P')$ can then be written as:

$$\text{AMSE}(P') = \sum p_j \cdot (1-p_j) / m$$

For $k=2$ we can rewrite this as:

$$\text{AMSE}(P') = 0.5/m - (p_1 - 0.5)^2/m - (p_2 - 0.5)^2/m$$

so

$$m \cdot \text{AMSE}(P') + (p_1 - 0.5)^2 + (p_2 - 0.5)^2 = 0.5$$

The surface is thus a cap of an elliptical parabola with vertex in the point $(0.5, 0.5, 1/2m)$. The height of this cap depends inversely on the total sample size.

For the predictor P^* , we get

$$\begin{aligned} \text{AMSE}(P^*) &= \sum f_j \cdot E(p^* - p_j)^2 = \\ &= \sum f_j \cdot (\sum m_j \cdot p_j \cdot (1-p_j) / m^2 + (p_j - \pi)^2) \\ &= \sum f_j \cdot p_j \cdot (1-p_j) / m + \sum f_j \cdot (p_j - \pi)^2 \end{aligned}$$

$$\text{where } \pi = E(p^*) = \sum f_j \cdot p_j$$

Thus, $\text{AMSE}(P^*)$ consists of two parts, measuring the variance and the bias respectively. We note that the variance part is always less or equal than $\text{AMSE}(P')$. For $k=2$, $\text{AMSE}(P^*)$ is equal to

$$\begin{aligned} \text{AMSE}(P^*) &= (f_1 \cdot (p_1 \cdot (1-p_1)) + (1-f_1) \cdot p_2 \cdot (1-p_2)) / m \\ &\quad + f_1 \cdot (1-f_1) \cdot (p_1 - p_2)^2 \end{aligned} \quad (4.1.1)$$

In fig 4.1.1 and 4.1.2 we plot $\text{AMSE}(P^*)$, for different values of p_1 and $p_2=1-p_1$ and $p_2=0.5-p_1$ respectively. The total sample size, m , is set equal to 30 in these and the following figures. We have also plotted $\text{AMSE}(P')$. These lines as well as the line $p_2=0.1-p_1$ are shown in fig 4.1.3.

The intersection of the graphs of $AMSE(P^*)$ and $AMSE(P')$ defines the interval where P^* performs better than P' . Due to a reduced bias term, this interval increases when f_1 moves away from 0.5. The length of the interval depends inversely on the total sample size.

Write the criterion $AMSE(P^*) - AMSE(P') > 0$ in the form

$$\delta_b = \frac{m \cdot \sum f_j \cdot (p_j - \pi)^2}{\sum (1 - f_j) \cdot p_j \cdot (1 - p_j)} > 1$$

where the index b refers to the binomial case. By making an appropriate change of coordinate system we will show that for $k=2$ relations of the kind $\delta_b=d$ defines an ellipse in the $p_1 \times p_2$ -space. Put

$$p_1 = x \cdot \cos\theta - y \cdot \sin\theta$$

$$p_2 = x \cdot \sin\theta + y \cdot \cos\theta$$

Substituting this into $\delta_b=d$, we get

$$A \cdot x^2 + B \cdot x \cdot y + C \cdot y^2 + D \cdot x + E \cdot y = 0 \quad (4.1.2)$$

where

$$A = (1+d/m_1) \cdot \cos^2\theta - 2 \cdot \cos\theta \cdot \sin\theta + (1+d/m_2) \cdot \sin^2\theta$$

$$B = -(d/m_1 - d/m_2) \cdot \sin 2\theta - 2 \cdot \cos 2\theta$$

$$C = (1+d/m_1) \cdot \sin^2\theta + 2 \cdot \cos\theta \cdot \sin\theta + (1+d/m_2) \cdot \cos^2\theta$$

$$D = -(d/m_1) \cdot \cos\theta - (d/m_2) \cdot \sin\theta$$

$$E = (d/m_1) \cdot \sin\theta - (d/m_2) \cdot \cos\theta$$

Putting $B=0$ is equivalent to

$$\cot 2\theta = (1/m_2 - 1/m_1) \cdot d/m = d \cdot (2 \cdot f_1 - 1) / 2 \cdot m \cdot f_1 \cdot (1 - f_1)$$

Solving for θ

$$\theta = \pi/4 - (1/2) \cdot \arctan(d \cdot (2 \cdot f_1 - 1) / 2 \cdot m \cdot f_1 \cdot (1 - f_1))$$

$$\approx \pi/4 - d \cdot (2 \cdot f_1 - 1) / 4 \cdot m \cdot f_1 \cdot (1 - f_1) \quad (4.1.3)$$

It is easily seen that, since $d > 0$, both A and B are larger than zero for all values of θ . This proves that the relation $\delta_b = d$ defines an ellipse. For $f_1 = 0.5$, $\theta = \pi/4$. Substituting this value into (4.1.2), we arrive at

$$d \cdot x^2/2 - d \cdot x \cdot 2^{-0.5} + (m+d) \cdot y^2/2 = 0$$

Completing squares and rewriting in the standard formula for an ellipse we get

$$\frac{(x - 2^{-0.5})^2}{1/2} + \frac{y^2}{d/2 \cdot (m+d)} = 1 \quad (4.1.4)$$

The area inside the ellipse for $d=1$, defines together with the requirement $0 \leq p_1, p_2 \leq 1$, the region where P^* is preferred to P' . The eccentricity of this ellipse is

$$\sqrt{1 - \frac{1}{(m+1)^2}}$$

and thus it becomes flatter for large m , and because the major axis is constant this indicates that the region where P^* is preferred becomes smaller. By looking at (4.1.3) and (4.1.4), we see that the major axis coincide with the $p_1=p_2$ -line if $f_1=0.5$. For $f_1 \neq 0.5$, the major axis is tilted off this line. The magnitude of this effect is inversely related to m . For a total sample size of $m=30$, the region was plotted for two cases, $f_1=15/30$ and $f_1=5/30$. The result is shown in fig 4.1.3, where it is seen that the region is larger for $f_1 = 5/30$. By inspecting (4.1.1), we see that this is no accident, since as a function of f_1 , $AMSE(P^*)$ reaches its maximum for $f_1=0.5$ and $AMSE(P')$ don't depend on f_1 .

4.2 The suggested predictor P_b^- .

Now, ideally we would use the predictor P' whenever $\delta_b > 1$ and otherwise the predictor P^* . If this prior knowledge is not available we must estimate the criterion. For this purpose we will use the maximum likelihood-method.

It is a well known fact that the M.L.-estimators of p_j and π under the unrestricted model are respectively:

$$ml(\pi) = \sum f_j \cdot p_j' = p^* \quad \text{and} \quad ml(p_j) = p_j' = x_j/n_j$$

and substituting these into δ_b , we obtain the estimated criterion:

$$\delta_b' = \frac{n \cdot \sum f_j \cdot (p_j' - p^*)^2}{\sum (1-f_j) \cdot p_j' \cdot (1-p_j')} > 1$$

We note here that δ_b' is equivalent to a statistic proposed by Goodman (1964) as a competitor to the chi-squared test for

homogeneity. In the literature it is also known as the Wald statistic.

Define a prediction rule

Use P' if $\delta_b' > 1$

Use P^* if $\delta_b' \leq 1$

and denote this predictor P_b^{\sim} . Putting $PR_b = P(\delta_b' > 1)$, the average mean squared error of P_b^{\sim} is:

$$AMSE(P_b^{\sim}) = PR_b \cdot AMSE(P') + (1 - PR_b) \cdot AMSE(P^*) =$$

$$PR_b \cdot \sum p_j \cdot (1 - p_j) / n +$$

$$(1 - PR_b) \cdot (\sum f_j \cdot p_j \cdot (1 - p_j) +$$

$$\sum f_j \cdot (p_j - \pi)^2)$$

Of course, we would like PR_b to be as large as possible whenever $\delta_b \leq 1$ and as small as possible when $\delta_b > 1$. Further, to calculate $AMSE(P_b^{\sim})$ for different situations we must be able to compute the value of PR_b . This was done through simulations for

the two-population case, first for $f_1=15/30$ and second for $f_1=5/30$. The result is presented in fig 4.2.1 and 4.2.2 where $AMSE(P')$ and $AMSE(P^*)$ are also plotted.

4.3 Comparison of P_b and predictors based on Pearson's chi-squared test.

We now proceed to show that the statistic δ_b' has a close connection with the Pearson chi-square statistic used for testing the equality of k binomial probabilities:

$$x^2 = \frac{n \cdot \sum f_j \cdot (p_j' - p^*)^2}{p^* \cdot (1 - p^*)}$$

The difference between the two statistics lies in their denominators. Expressing δ_b' in terms of x^2 we get

$$\delta_b' = \frac{p^* \cdot (1 - p^*)}{\sum (1 - f_j) \cdot p_j' \cdot (1 - p_j')} x^2 =$$

$$= \frac{\sum f_j \cdot (p_j' - p^*)^2 + \sum f_j \cdot p_j' \cdot (1 - p_j')}{\sum (1 - f_j) \cdot p_j' \cdot (1 - p_j')} x^2 =$$

$$\frac{\sum f_j \cdot p_j' \cdot (1 - p_j')}{\sum (1 - f_j) p_j' \cdot (1 - p_j)} = \frac{x^2}{1 - x^2/n}$$

To give a numerical example to show that δ_b' generally is not a function of x^2 , we pick two values of (p_1, p_2) , for which $x^2=1$ and compute the value of δ_b' in both cases. For the values $(0.20, 0.0629)$ and $(0.20, 0.4400)$ $x^2=1$, if $f_1=5/30$. In the first case $\delta_b'=0.55$, while in the second case we get $\delta_b'=1.38$.

In the case of equal sampling proportions, $f_j=1/k$ $j=1, 2, \dots, k$, we however get

$$\delta_b' = \frac{x^2/(k-1)}{1 - x^2/n}$$

Thus, we can state the criterion for choosing between P' and P^* equivalently in terms of χ^2 :

$$\delta_{p'} > 1 \quad \Leftrightarrow \quad \chi^2 > n \cdot (k-1) / (n+k-1) \approx k-1$$

For different values of k , using the criterion $\delta_{p'}$ corresponds to adopting a χ^2 - test at the following approximate levels:

k	α
2	0.32
3	0.38
4	0.40
5	0.41
>30	0.50

Let P_α be the predictor defined by the following rule:

Use P' if $X^2 > c_{1-\alpha}$

Use P^* if $X^2 < c_{1-\alpha}$

where $c_{1-\alpha}$ is the upper $(1-\alpha)$ -percentile in a X^2 -distribution with $k-1$ degrees of freedom and let $PR_\alpha = P(X^2 > c_{1-\alpha})$. The AMSE of P_α can be written as

$$AMSE(P_\alpha) = PR_\alpha \cdot AMSE(P') + (1-PR_\alpha) \cdot AMSE(P^*)$$

Next, we will compare $AMSE(P_{b^-})$ and $AMSE(P_\alpha)$ for the two-population case. Referring to the discussion above, it is clear that for equal sampling proportions ($f_1=0.5$), the criterion $\delta_{b^-} > 1$ is equivalent to $X^2 > n/(n+1)$. The latter critical value is for reasonably large n approximately equal to 1 and it corresponds to a test on the approximate level of $\alpha=0.32$.

The difference between the two AMSE:s can be written as

$$AMSE(P_{b^-}) - AMSE(P_\alpha) = (PR_b - PR_\alpha) \cdot (AMSE(P') - AMSE(P^*))$$

and we conclude that for equal sampling proportions we have

$AMSE(P_{b\sim}) - AMSE(P_{\alpha}) = 0$ if $AMSE(P') = AMSE(P^*)$ or

$$PR_b = PR_{\alpha}$$

> 0 if $AMSE(P') > AMSE(P^*)$ and $\alpha < 0.32$
or $AMSE(P') < AMSE(P^*)$ and $\alpha > 0.32$

< 0 if $AMSE(P') > AMSE(P^*)$ and $\alpha > 0.32$
or $AMSE(P') < AMSE(P^*)$ and $\alpha < 0.32$

To illustrate this and also study the effect of nonequal sampling proportions, $AMSE(P_{b\sim})$ and $AMSE(P_{\alpha})$ were computed through simulations. Two critical values for the χ^2 -test were considered, 3.84 and 0.45, corresponding to the approximate levels 0.05 and 0.50 respectively. Two sampling proportions were chosen, $f_1=15/30$ and $f_1=5/30$. The results are summarized in fig 4.3.1 - 4.3.6.

As all curves intersect at the points where $AMSE(P') = AMSE(P^*)$, fig 4.3.1-4.3.6 illustrate how the region where P^* is preferred to P' is larger for $f_1=5/30$ than for $f_1=15/30$. This is the result of a decrease in the average squared bias for P^* (see (4.1.1) page 15).

We can also conclude that a traditional strategy involving a low-level test performs well if p_1 and p_2 are close, while the $\delta_{b'}$ -criterion works better elsewhere. As noted earlier, the

effect of increasing m , would be to reduce the region where P^* is preferred to P' . Thus, for larger m , the δ_b' -criterion would be better for a yet larger region.

In fig 4.3.4 and 4.3.6, we also note the non-symmetry for $AMSE(P_{0.05})$, when we have non-equal sampling proportions. $AMSE(P_{0.05})$ is larger for small values of p_1 than for large values of p_1 .

Fig 4.3.6 shows that $AMSE(P_{0.50})$ actually is smaller than $AMSE(P_b')$ on the interval where $AMSE(P^*) < AMSE(P')$ if $f_1 = 5/30$, for the small values of p_1 and p_2 covered in this figure.

4.4 Comparing $P_{\tilde{p}}$ with predictors based on the Akaike-criterion.

The Akaike-criterion for choosing between two models, amounts to comparing the quantities

$$A_1 = l_1 + q_1 \quad \text{and}$$

$$A_2 = l_2 + q_2$$

where l_1 and l_2 are the maximized log-likelihood functions and q_1 and q_2 the number of estimated parameters of the models.

For our purposes we let l_1 be the maximized log-likelihood function under the hypothesis of homogeneity

$$l_1 = \Sigma (x_{1j} \cdot \ln(p^*) + (n_j - x_{1j}) \cdot \ln(1-p^*))$$

and under the global alternative hypothesis we get

$$l_2 = \Sigma (x_{1j} \cdot \ln(p_j') + (n_j - x_{1j}) \cdot \ln(1-p_j'))$$

The Akaike-criterion states that we should use the predictor P' if

$$A_2 - A_1 > 0 \iff l_2 - l_1 > q_2 - q_1$$

Equivalently we may express this in terms the likelihood functions, L_1 and L_2 :

$$\ln(L_2/L_1) > q_2 - q_1$$

For the two-population case $q_2 - q_1 = 1$, so we obtain

$$\ln(L_2/L_1) > 1$$

Since $2 \cdot \ln(L_2/L_1)$ has an approximate chi-square null distribution with one degree of freedom, this criterion is approximately equivalent to adopting a likelihood ratio test on the level of 0.16.

Let P_A be the predictor defined by the rule

$$\text{Use } P' \text{ if } \ln(L_2/L_1) > 1$$

$$\text{Use } P^* \text{ if } \ln(L_2/L_1) < 1$$

In fig 4.4.1 - 4.4.4 $AMSE(P_{\tilde{b}})$ is compared with $AMSE(P_A)$.

We see that P_A performs slightly better in the region where $\delta_b < 1$, corresponding to values of p_1 and p_2 quite close, while p_b^{\sim} is better elsewhere.

4.5 Effects of different sampling proportions in the old and the new sample.

In so far we have assumed that the sampling proportions in the old sample were identical to the proportions in the new sample, for which we wanted to make predictions. This assumption simplified the computations for $AMSE(P')$ and $AMSE(P^*)$. In this section we shall give a brief indication to what happens if this assumption is not fulfilled. Let

$e_j = m_j/m$, i.e. the proportion of obs. at $Z=j$
in the old sample, $j=1,2,\dots,k$

$f_j = n_j/n$, i.e. the proportion of obs. at $Z=j$
in the new sample, $j=1,2,\dots,k$

The $AMSE$:s are defined by averaging over the new sample as usual

$$AMSE(P') = \sum f_j \cdot MSE(p_j')$$

$$AMSE(P^*) = \sum f_j \cdot MSE(p_j^*)$$

Evaluating $MSE(p_j')$ and $MSE(p_j^*)$ for the proportions e_j , we get

$$AMSE(P') = \sum (f_j/m \cdot e_j) p_j \cdot (1-p_j)$$

$$AMSE(P^*) = \sum f_j \cdot ((\sum e_j \cdot p_j \cdot (1-p_j)) / m + (p_j - \pi)^2) =$$

$$\sum (e_j/m) \cdot p_j \cdot (1-p_j) + \sum f_j \cdot (p_j - \pi)^2$$

$$\text{where } \pi = \sum e_j \cdot p_j$$

Studying $AMSE(P^*)$ for the case where $k=2$, we obtain

$$AMSE(P^*) = \sum (e_j/m) \cdot p_j \cdot (1-p_j) +$$

$$(e_1^2 + (1 - 2 \cdot e_1) \cdot f_1) \cdot (p_1 - p_2)^2$$

Comparing $AMSE(P^*)$ for $e_1=f_1$ and for $e_1 \neq f_1$

$$AMSE(P^*)_{e_1=f_1} - AMSE(P^*)_{e_1 \neq f_1} = (1 - 2 \cdot e_1) \cdot (e_1 - f_1) \cdot (p_1 - p_2)^2$$

$$e_1=f_1 \quad e_1 \neq f_1$$

e.i. if $p_1 \neq p_2$ we are better off if we try to predict for the new sample, where $f_1 \neq e_1$ if $e_1 < 0.5$ and $f_1 < e_1$ or $e_1 > 0.5$ and $f_1 > e_1$. As is seen from fig 4.5.1, in the majority of cases we are however worse off. Note that we can't interpret fig 4.5.1

as indicating the effect of e_1 for a given f_1 . For example, if $f_1=0.9$, we can't say that $AMSE(P^*)$ is smaller for $e_1=0.6$, say, than for $e_1=0.9$. For the latter problem we can minimize

$$AMSE(P^*) = (1/m) \cdot \sum e_j \cdot p_j \cdot (1-p_j) + (e_1^2 + (1-2 \cdot e_1) \cdot f_1) \cdot (p_1-p_2)^2$$

with respect to e_1 . Taking derivative we obtain

$$\begin{aligned} d/de_1(AMSE(P^*)) &= p_1 \cdot (1-p_1)/m + p_2 \cdot (1-p_2)/m + \\ &+ 2 \cdot (e_1-f_1) \cdot (p_1-p_2)^2 \end{aligned}$$

Restricting attention to the case $p_1 \neq p_2$, for $p_1=p_2$ $AMSE(P^*)$ don't depend on either e_1 or f_1 , we put this equal to zero and solve for e_1 .

$$e_1 = f_1 + (p_2 \cdot (1-p_2) - p_1 \cdot (1-p_1)) / 2 \cdot m \cdot (p_1-p_2)^2$$

As the second derivative is equal to $4 \cdot (p_1-p_2)^2 > 0$, the solution is a minimum. It is seen that $e_1=f_1$ if the variances are equal in the two populations. On the other hand, if the variance $p_1 \cdot (1-p_1)$ is large compared to $p_2 \cdot (1-p_2)$, the minimal e_1 is smaller than f_1 .

Study the inequality

$$\begin{aligned} \text{AMSE}(P') - \text{AMSE}(P') &= (1 - f_1/e_1) \cdot p_1 \cdot (1-p_1) + \\ &\quad e_1=f_1 \quad e_1 \neq f_1 \\ &\quad (1 - (1-f_1)/(1-e_1)) \cdot p_2 \cdot (1-p_2) = \\ &= > 0 \end{aligned}$$

For $e_1 \neq f_1$ we get two cases

$$\begin{aligned} 1. \quad e_1 > f_1 \quad \text{AMSE}(P') - \text{AMSE}(P') > 0 \quad <=> \\ &\quad e_1=f_1 \quad e_1 \neq f_1 \end{aligned}$$

$$e_1 < \frac{p_1 \cdot (1 - p_1)}{p_1 \cdot (1 - p_1) + p_2 \cdot (1 - p_2)}$$

$$\begin{aligned} 2. \quad e_1 < f_1 \quad \text{AMSE}(P') - \text{AMSE}(P') > 0 \quad <=> \\ &\quad e_1=f_1 \quad e_1 \neq f_1 \end{aligned}$$

$$e_1 > \frac{p_1 \cdot (1 - p_1)}{p_1 \cdot (1 - p_1) + p_2 \cdot (1 - p_2)}$$

In fig 4.5.2 - 4.5.4, the regions in the (p_1, p_2) -space where

$$AMSE(P') - AMSE(P^*) > 0$$

$$e_1 = f_1 \quad e_1 \neq f_1$$

are shown, for values of e_1 corresponding to 0.4, 0.5 and 0.6. We see that for $e_1=0.5$, the inequality is satisfied for exactly half the space, both for $f_1 > e_1$ and $f_1 < e_1$. For $e_1=0.4$ it is satisfied for the larger region if $f_1 < 0.2$ and for $e_1=0.6$ if $f_1 > 0.6$.

It is clear that $AMSE(P_b^-)$, will be effected if $e_1 \neq f_1$. This effect should depend on the directions of the changes in $AMSE(P')$ and $AMSE(P^*)$. E.i. if both $AMSE(P')$ and $AMSE(P^*)$ get larger, then $AMSE(P_b^-)$ gets larger. This issue will not be discussed further.

4.6 The multinomial case.

In the preceding sections, we have assumed that the sample sizes for the different levels of Z , were fixed in advance, both in the old and the new sample. E.i. we were dealing with independent binomial sampling.

In this section we shall see that not much is changed for the case where the sample sizes for different levels of Z are random variables. We will assume that the total sample size is fixed, e.i. multinomial sampling. As before we have the two prediction rules P' and P^* . We define

$$AMSE(P') = \Sigma p_{.j} \cdot E(p_j' - p_j)^2 \quad \text{and}$$

$$AMSE(P^*) = \Sigma p_{.j} \cdot E(p^* - p_j)^2$$

where $p_{.j} = P(Z=j)$ in the population where we want to make predictions. Note that we keep the old notations for p_j' and p^* , but that they now have different distributions. After some computations we get

$$AMSE(P^*) = \Sigma p_{.j} \cdot p_j(1-p_j) + (m+1) \cdot \Sigma p_{.j} \cdot (p_j - p_{1.})^2/m$$

where $p_{1.} = P(Y=1)$. For $AMSE(P')$ we rely on an approximation for computing $V(p_j')$ (the deltha-method, see appendix):

$$V(p_j') \approx p_j \cdot (1-p_j) / m \cdot p_{.j}$$

so

$$AMSE(P') = \Sigma p_j \cdot (1-p_j) / m$$

Now, we would prefer the prediction rule P' if

$$\Sigma p_{.j} \cdot p_j \cdot (1-p_j) / m + (m+1) \Sigma p_{.j} \cdot (p_j - p_{1.})^2 / m - \Sigma p_j \cdot (1-p_j) / m > 0$$

e.i. if

$$\delta_m = \frac{m \cdot \Sigma p_{.j} \cdot (p_j - p_{1.})^2}{\Sigma (1 - p_{.j}) \cdot p_j \cdot (1 - p_j)} > 1$$

So δ_m is practically equal to δ_b . To estimate δ_m , we insert the M-L-estimates of the parameters (see appendix), and obtain:

$$\delta_m' = \frac{m \cdot \sum p_{.j}' \cdot (p_j - p^*)^2}{\sum (1 - p_{.j}') \cdot p_{.j}' \cdot (1 - p_{.j}')} > 1$$

where $p_{.j}' = x_{.j}/m$, the proportion of observations at $Z=j$, in the old sample. Applying the prediction rule

Use P' if $\delta_m' > 1$

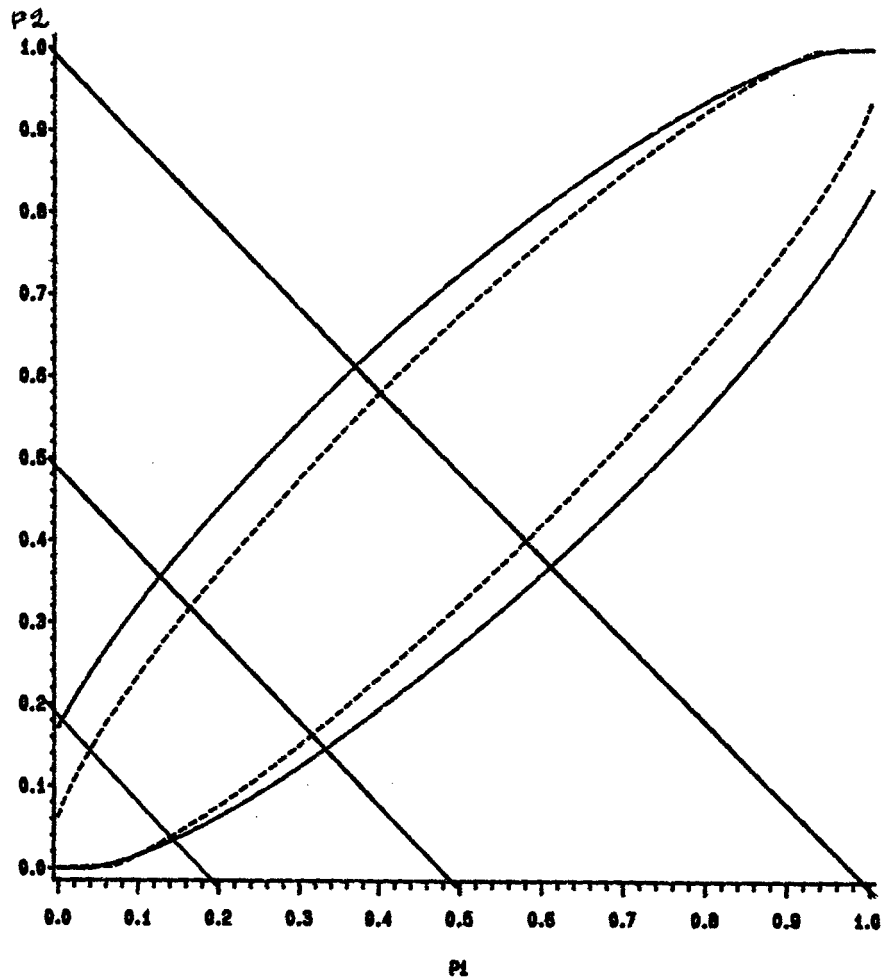
Use P^* if $\delta_m' \leq 1$

and call this rule P_m' . Putting $PR(\delta_m' > 1)$, The AMSE of P_m' can be written as

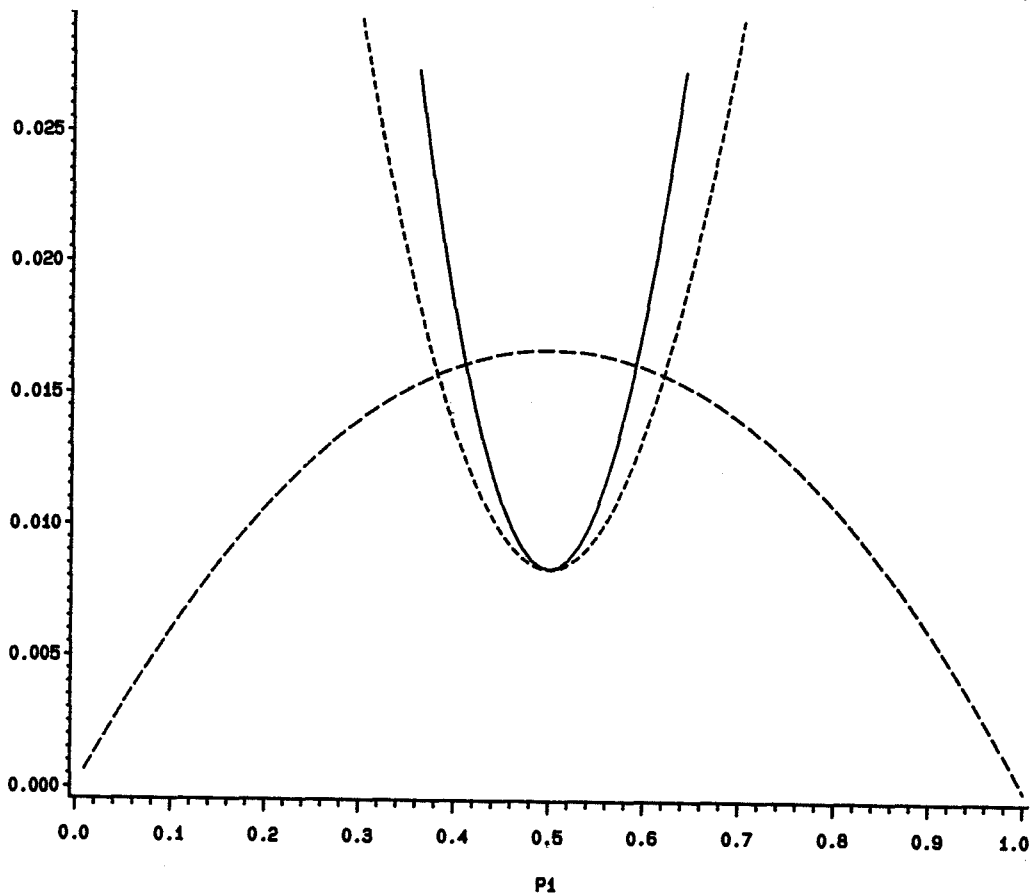
$$AMSE(P_m') = PR_m \cdot AMSE(P') + (1 - PR_m) \cdot AMSE(P^*)$$

As in the binomial case this is an average of $AMSE(P')$ and $AMSE(P^*)$. Thus P_m' can be expected to perform well over large regions of the parameter space.

Fig 4.1.3 Values of p_1 and p_2 where $\delta_b = 1$.

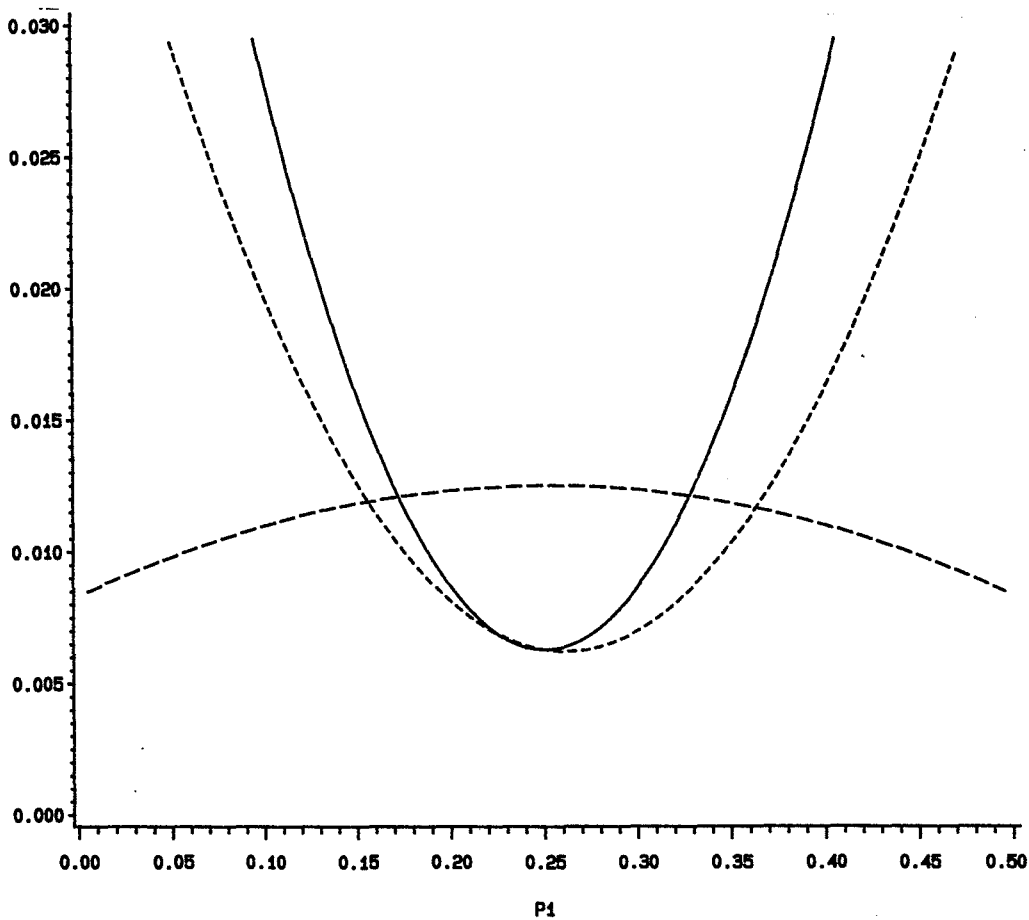


$f_1 = 5/30$: — , $f_1 = 15/30$ --- . The lines $p_2 = 1 - p_1$, $p_2 = 0.5 - p_1$ and $p_2 = 0.2 - p_1$ are also indicated.



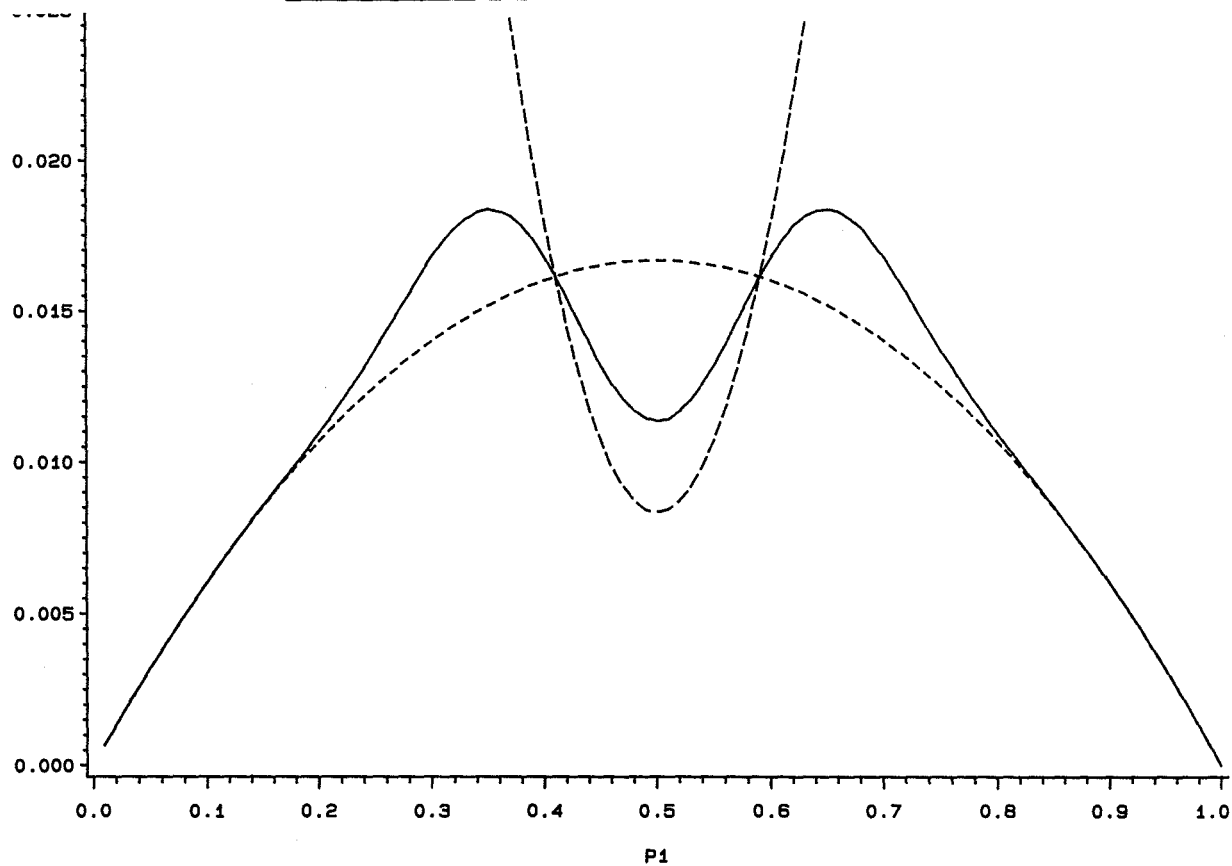
$AMSE(P')$:---, $AMSE(P^*)$ $f_1=5/30$:--- $AMSE(P^*)$ $f_1=15/30$:—

Fig 4.1.2 Comparison of $AMSE(P')$ and $AMSE(P^*)$ for $p_2=0.5-p_1$.



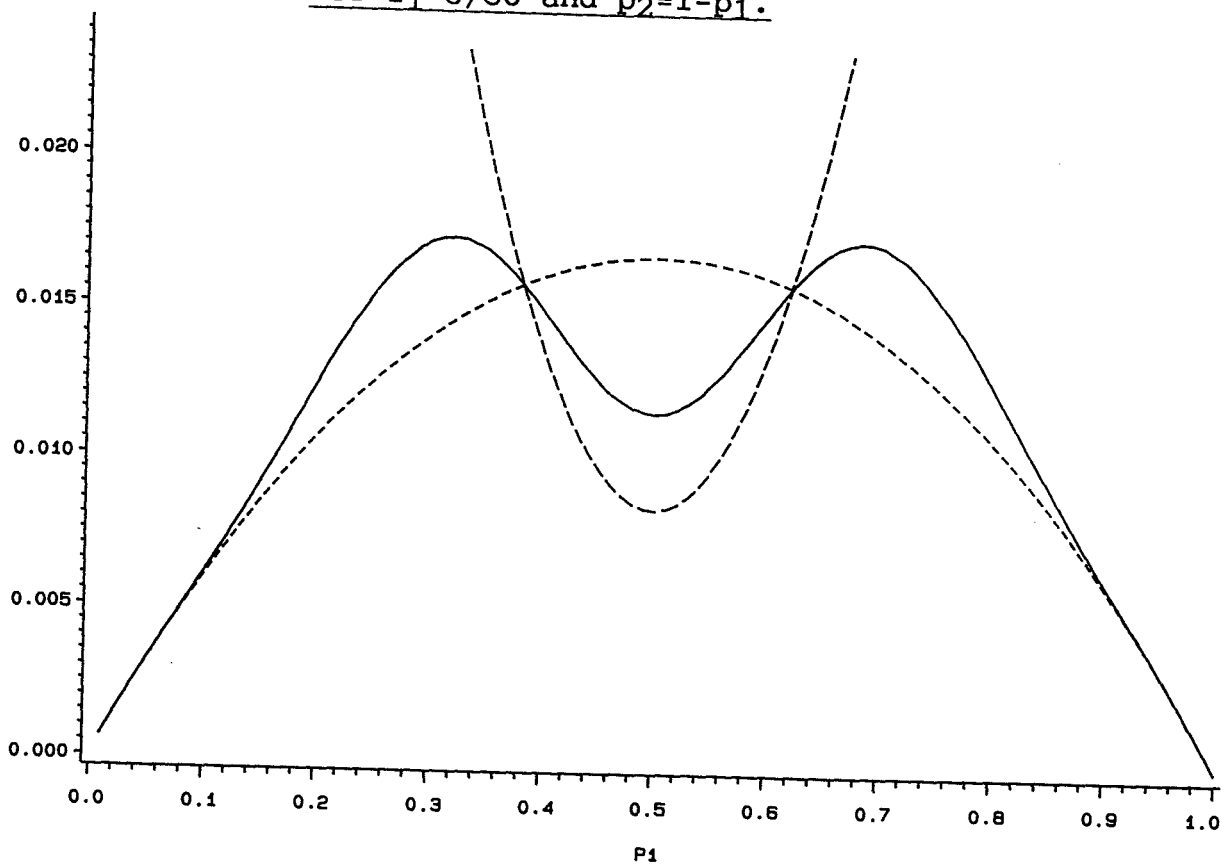
$AMSE(P')$:---, $AMSE(P^*)$ $f_1=5/30$:--- $AMSE(P^*)$ $f_1=15/30$:—

Fig 4.2.1 Comparison of $AMSE(P')$, $AMSE(P^*)$ and $AMSE(P_b^-)$.
for $f_1=15/30$ and $p_2=1-p_1$.



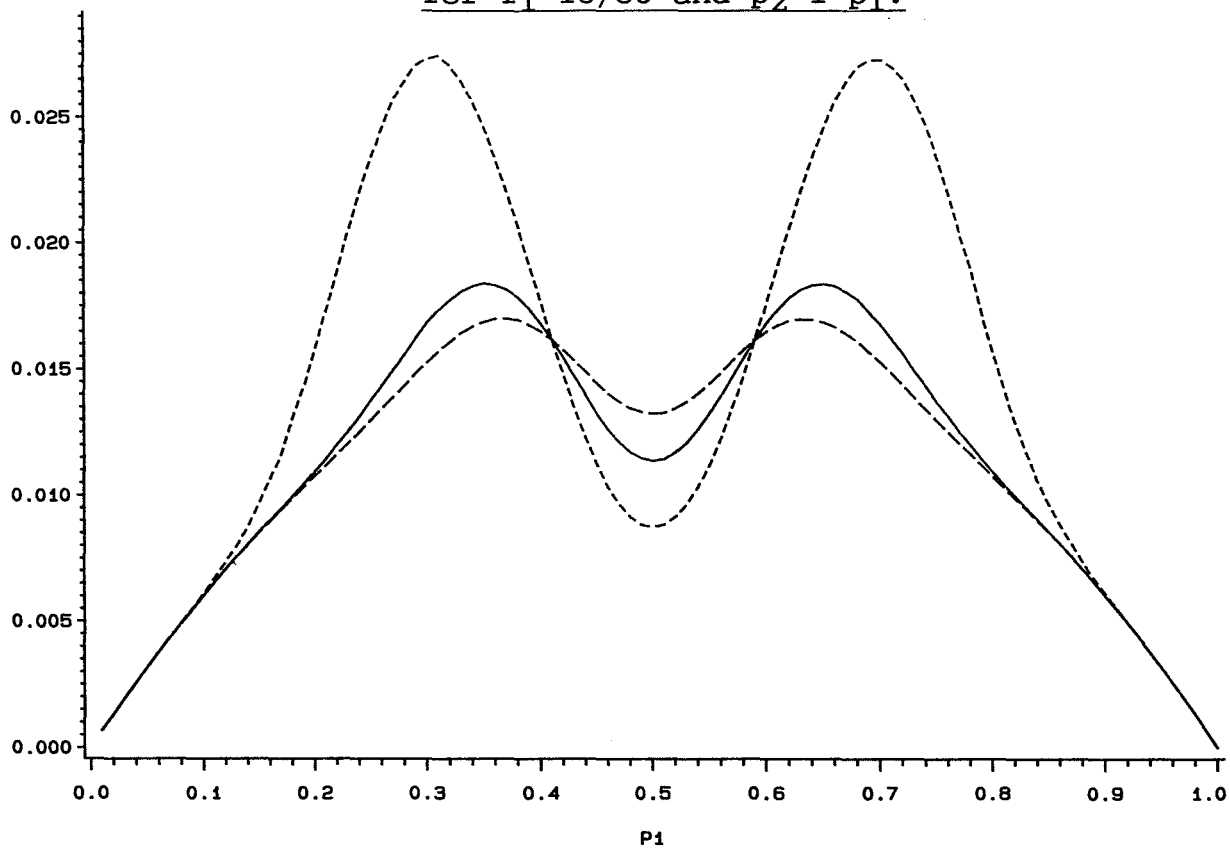
$AMSE(P')$: - . - . , $AMSE(P^*)$: - - - - , $AMSE(P_b^-)$: —

Fig 4.2.2 Comparison of $AMSE(P')$, $AMSE(P^*)$ and $AMSE(P_b^-)$
for $f_1=5/30$ and $p_2=1-p_1$.



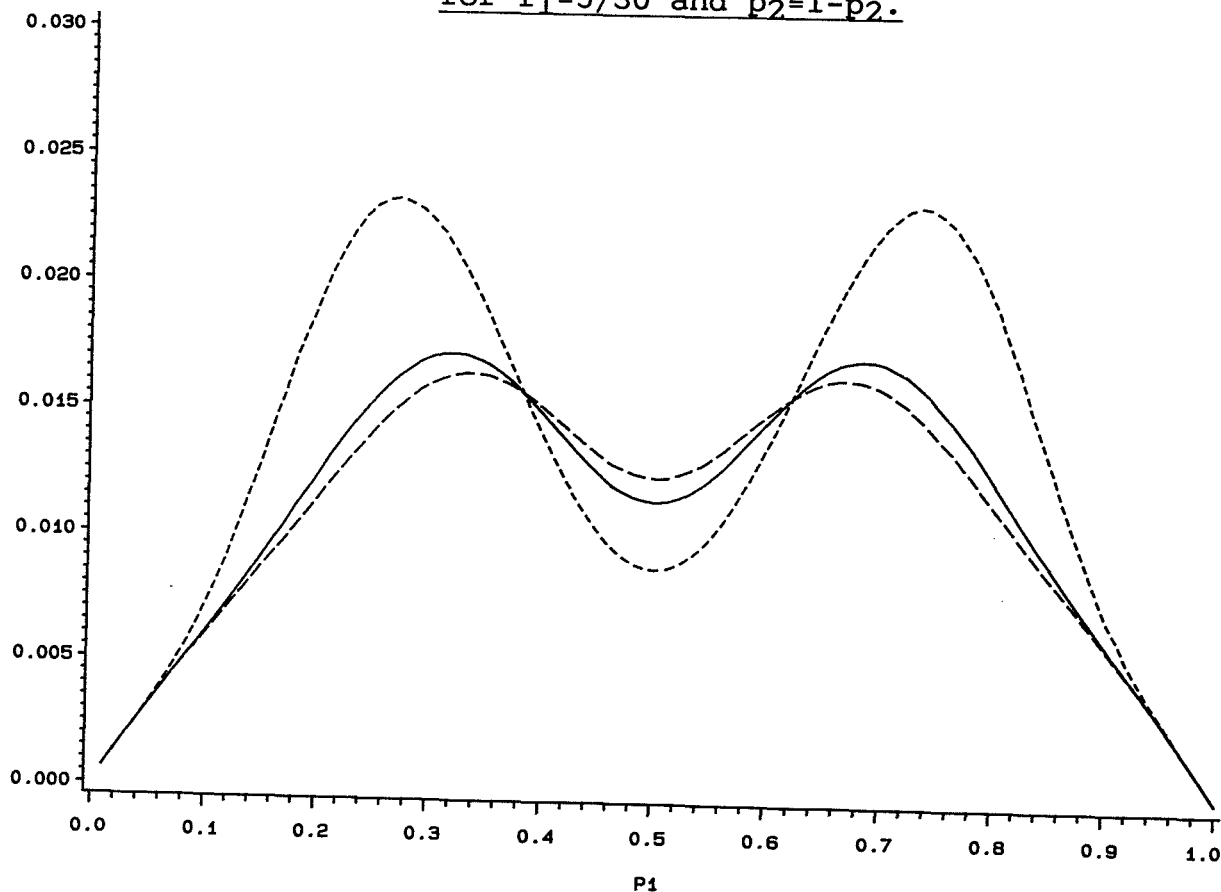
$AMSE(P')$: - . - . , $AMSE(P^*)$: - - - - , $AMSE(P_b^-)$: —

Fig 4.3.1 Comparison of $AMSE(P_{0.50})$, $AMSE(P_{0.05})$ and $AMSE(P_{b^{\sim}})$ for $f_1=15/30$ and $p_2=1-p_1$. 55



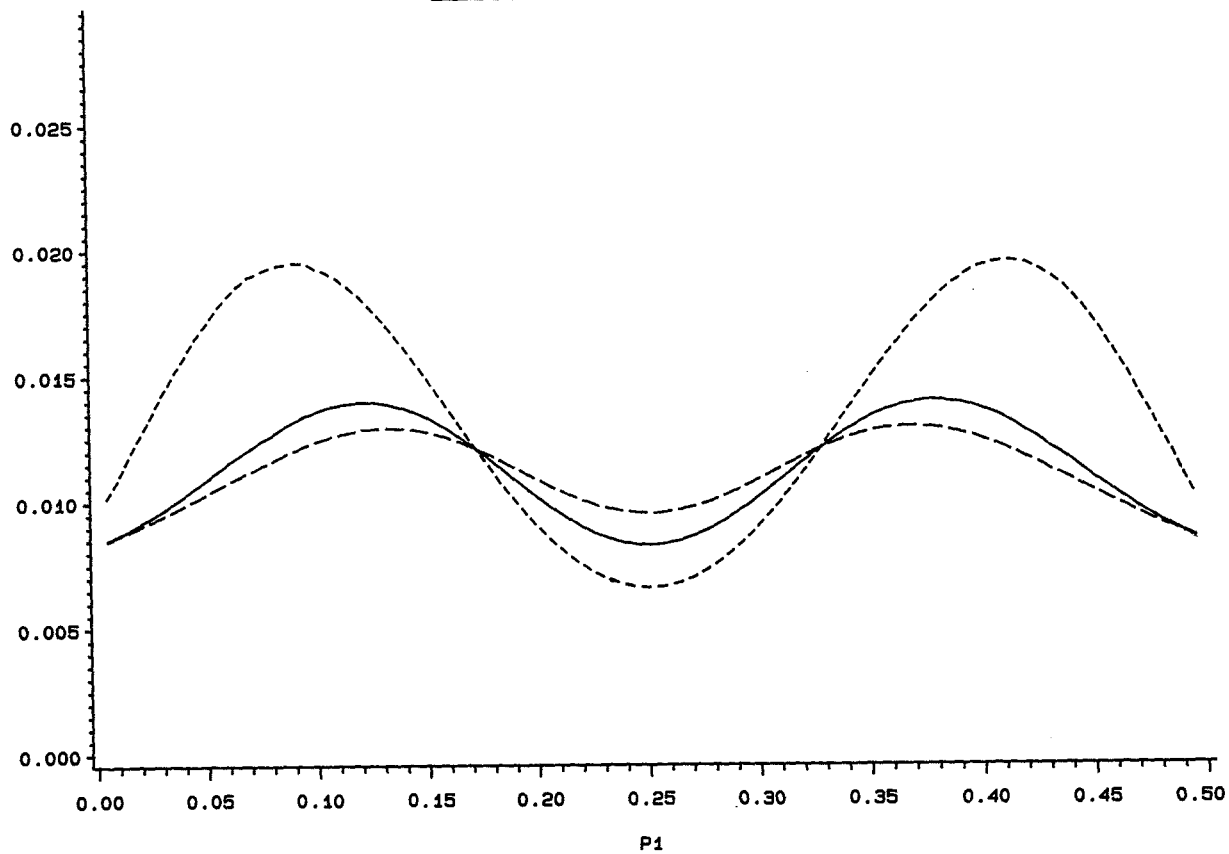
$AMSE(P_{0.50})$: - - - , $AMSE(P_{0.05})$: - . - , $AMSE(P_{b^{\sim}})$: —

Fig 4.3.2 Comparison of $AMSE(P_{0.50})$, $AMSE(P_{0.05})$ and $AMSE(P_{b^{\sim}})$ for $f_1=5/30$ and $p_2=1-p_2$.



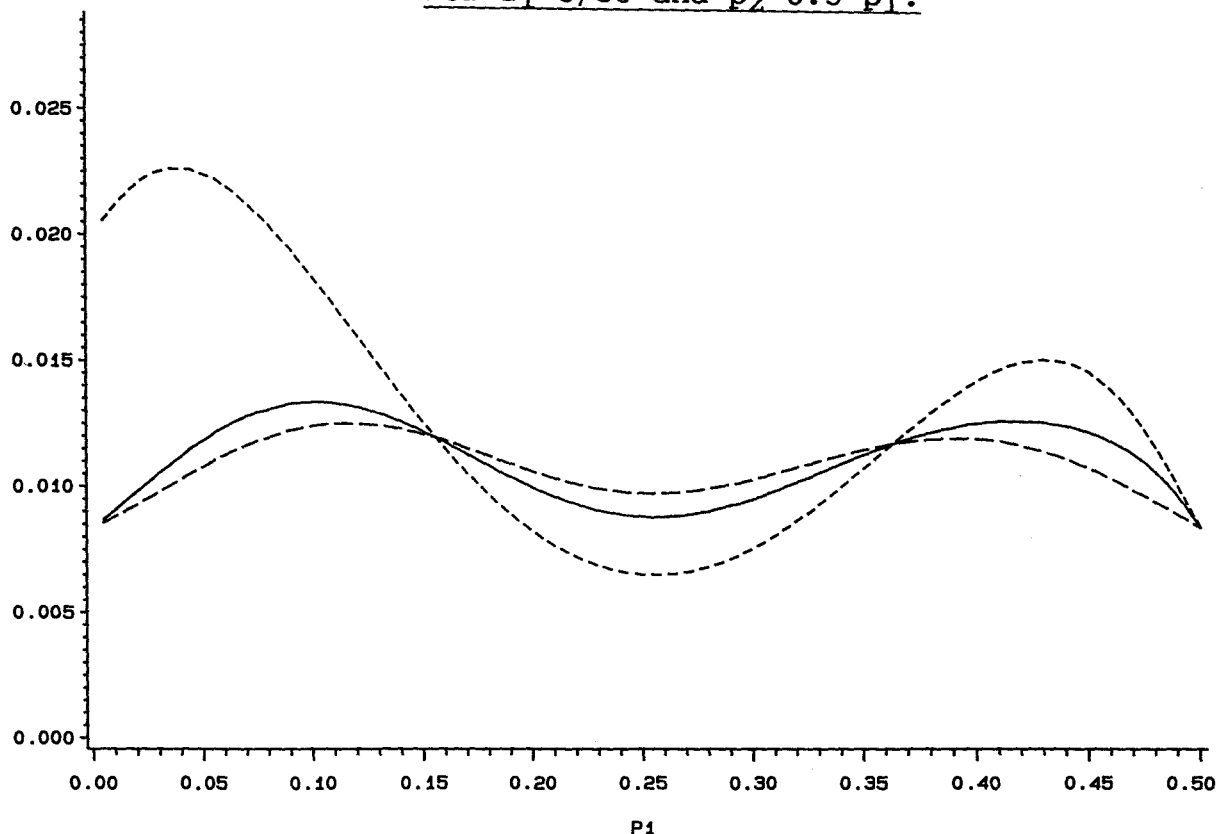
$AMSE(P_{0.50})$: - - - , $AMSE(P_{0.05})$: - . - , $AMSE(P_{b^{\sim}})$: —

Fig 4.3.3 Comparison of $AMSE(P_{0.50})$, $AMSE(P_{0.05})$ and $AMSE(P_{b^*})$ for $f_1=15/30$ and $p_2=0.5-p_1$.



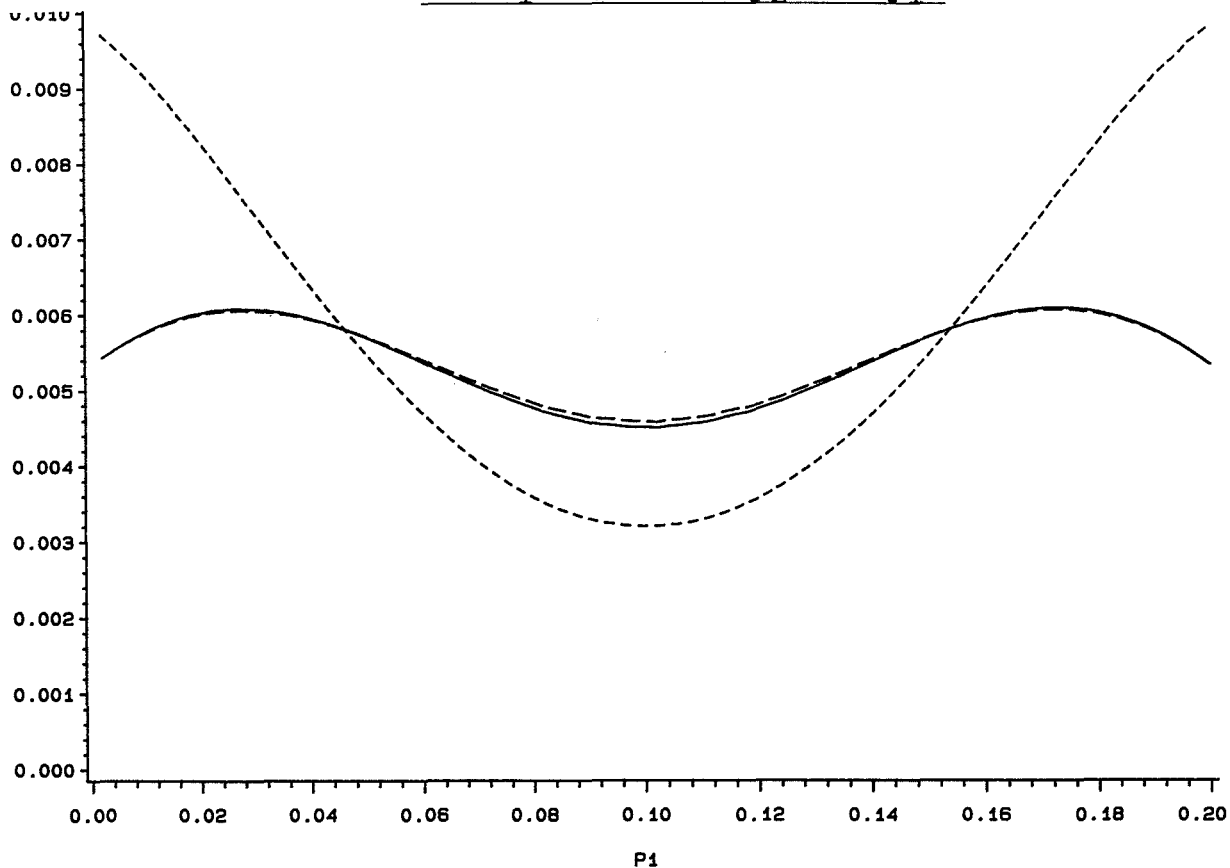
$AMSE(P_{0.50})$:--- , $AMSE(P_{0.05})$:-.- , $AMSE(P_{b^*})$:—

Fig 4.3.4 Comparison of $AMSE(P_{0.50})$, $AMSE(P_{0.05})$ and $AMSE(P_{b^*})$ for $f_1=5/30$ and $p_2=0.5-p_1$.



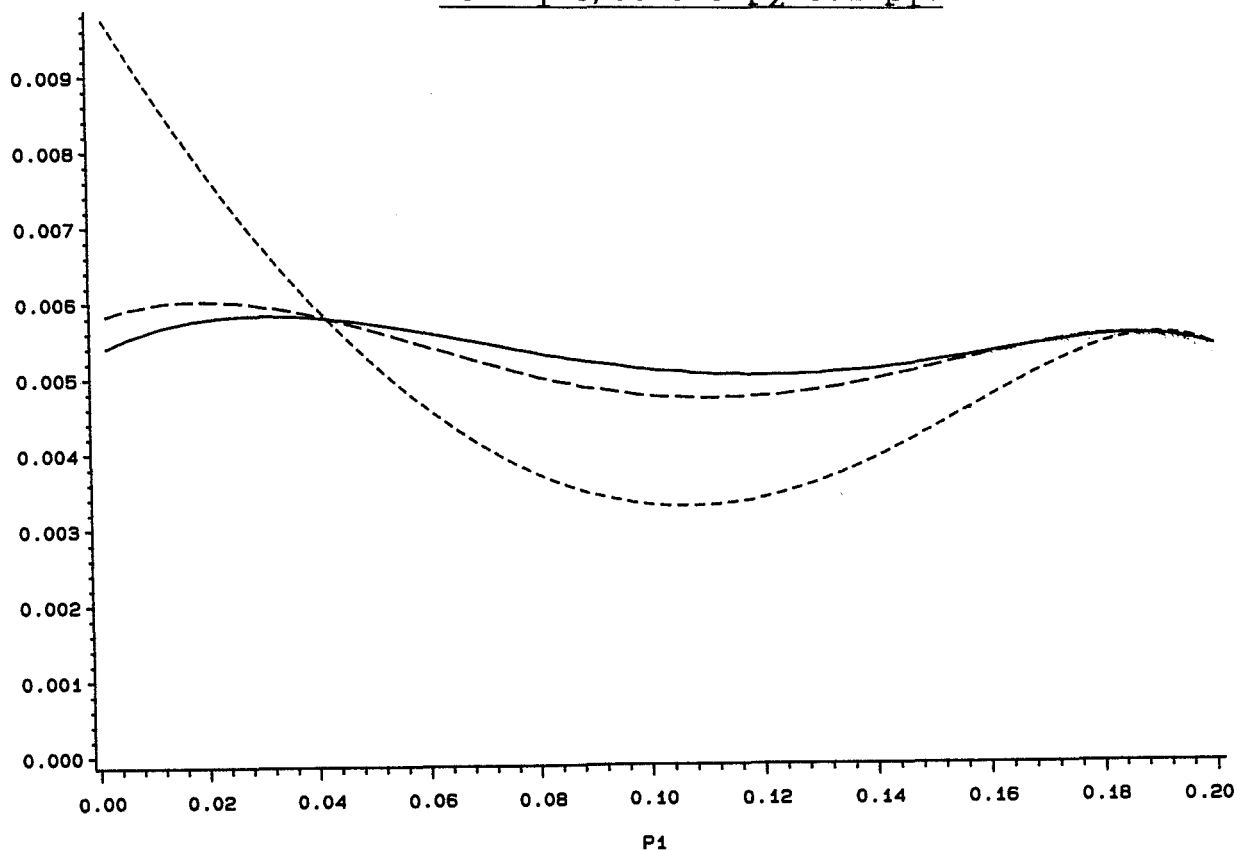
$AMSE(P_{0.50})$:--- , $AMSE(P_{0.05})$:-.- , $AMSE(P_{b^*})$:—

Fig 4.3.5 Comparison of AMSE($P_{0.50}$), AMSE($P_{0.05}$) and AMSE($P_{b^{\sim}}$)⁵⁷
for $f_1=15/30$ and $p_2=0.2-p_1$.



AMSE($P_{0.50}$):---, AMSE($P_{0.05}$):-.-, AMSE($P_{b^{\sim}}$):—

Fig 4.3.6 Comparison of AMSE($P_{0.50}$), AMSE($P_{0.05}$) and AMSE($P_{b^{\sim}}$)
for $f_1=5/30$ and $p_2=0.2-p_1$.



AMSE($P_{0.50}$):---, AMSE($P_{0.05}$):-.-, AMSE($P_{b^{\sim}}$):—

Fig 4.4.1 Comparison of $AMSE(P_A)$ and $AMSE(P_{b^-})$
for $f_1=15/30$, $p_2=0.5-p_1$.

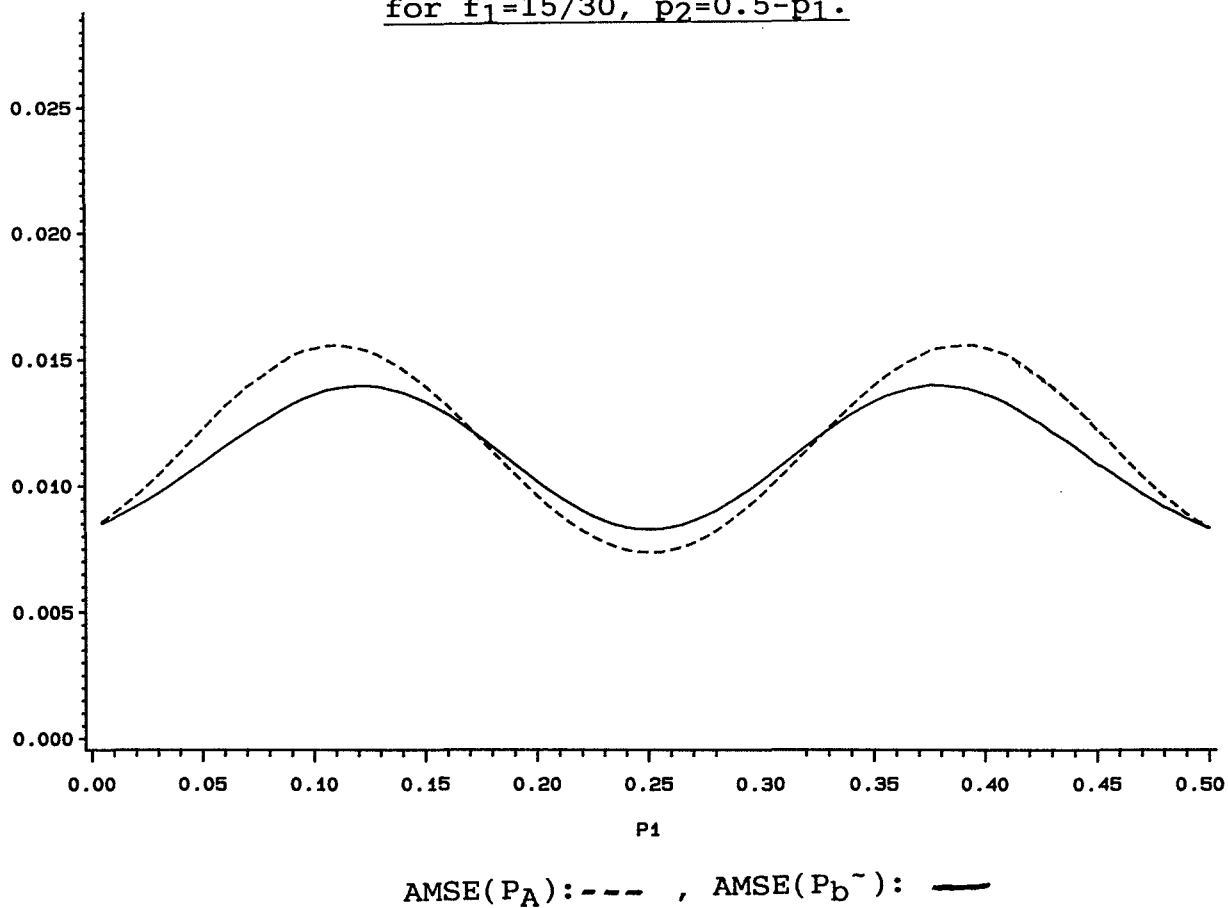


Fig 4.4.2 Comparison of $AMSE(P_A)$ and $AMSE(P_{b^-})$
for $f_1=5/30$, $p_2=0.5-p_1$.

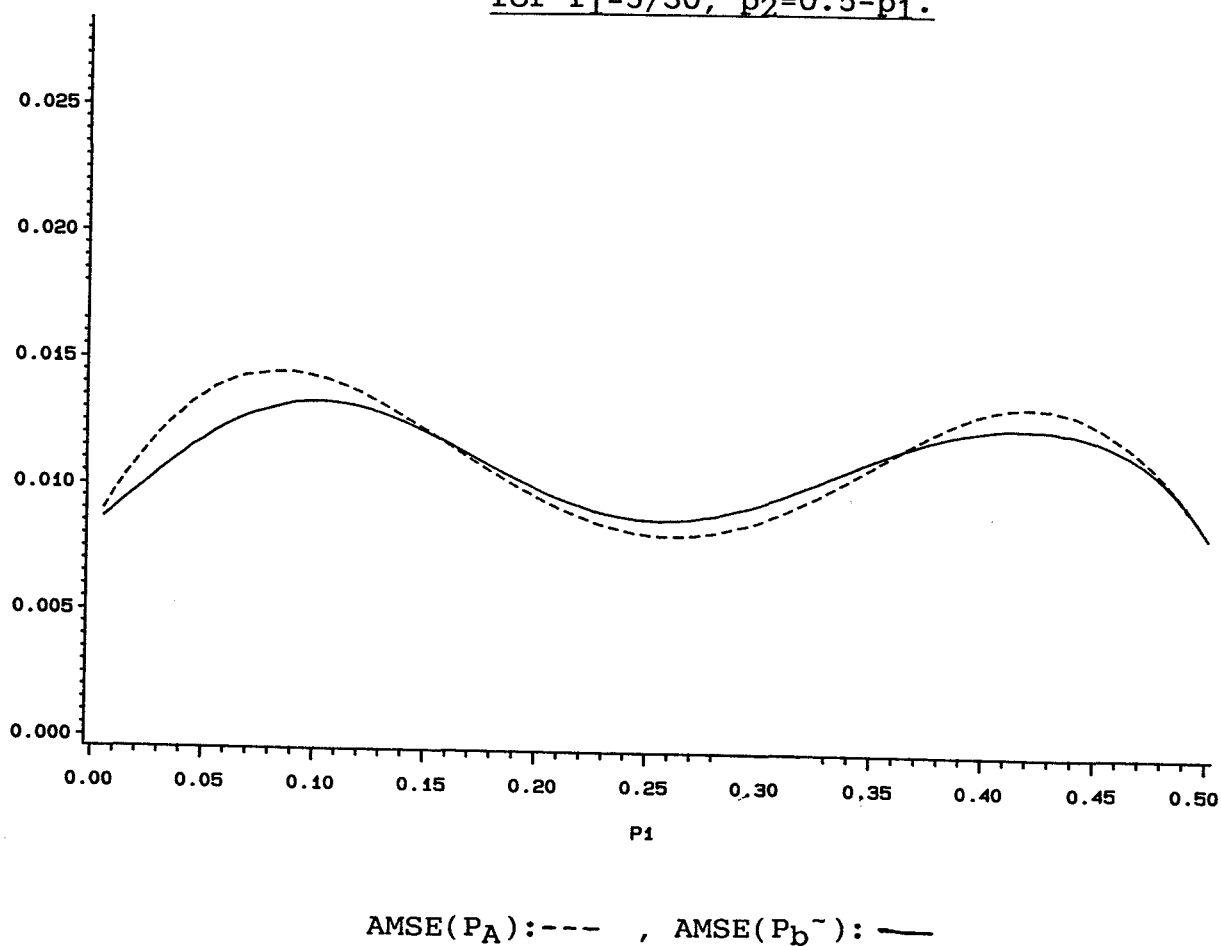


Fig 4.4.3 Comparison of $AMSE(P_A)$ and $AMSE(P_{b\sim})$
for $f_1=15/30, p_2=0.2-p_1$.

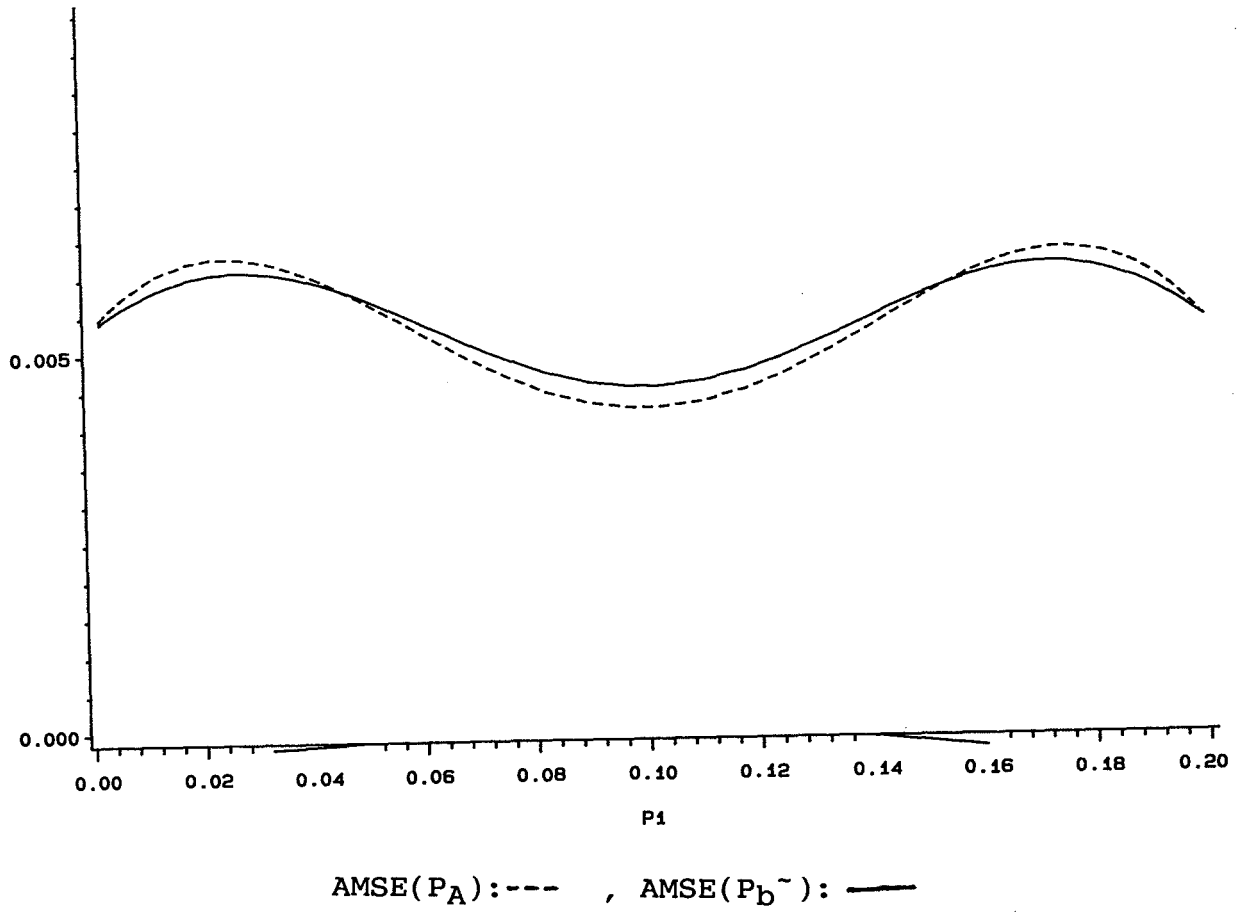


Fig 4.4.4 Comparison of $AMSE(P_A)$ and $AMSE(P_{b\sim})$
for $f_1=5/30, p_2=0.2-p_1$.

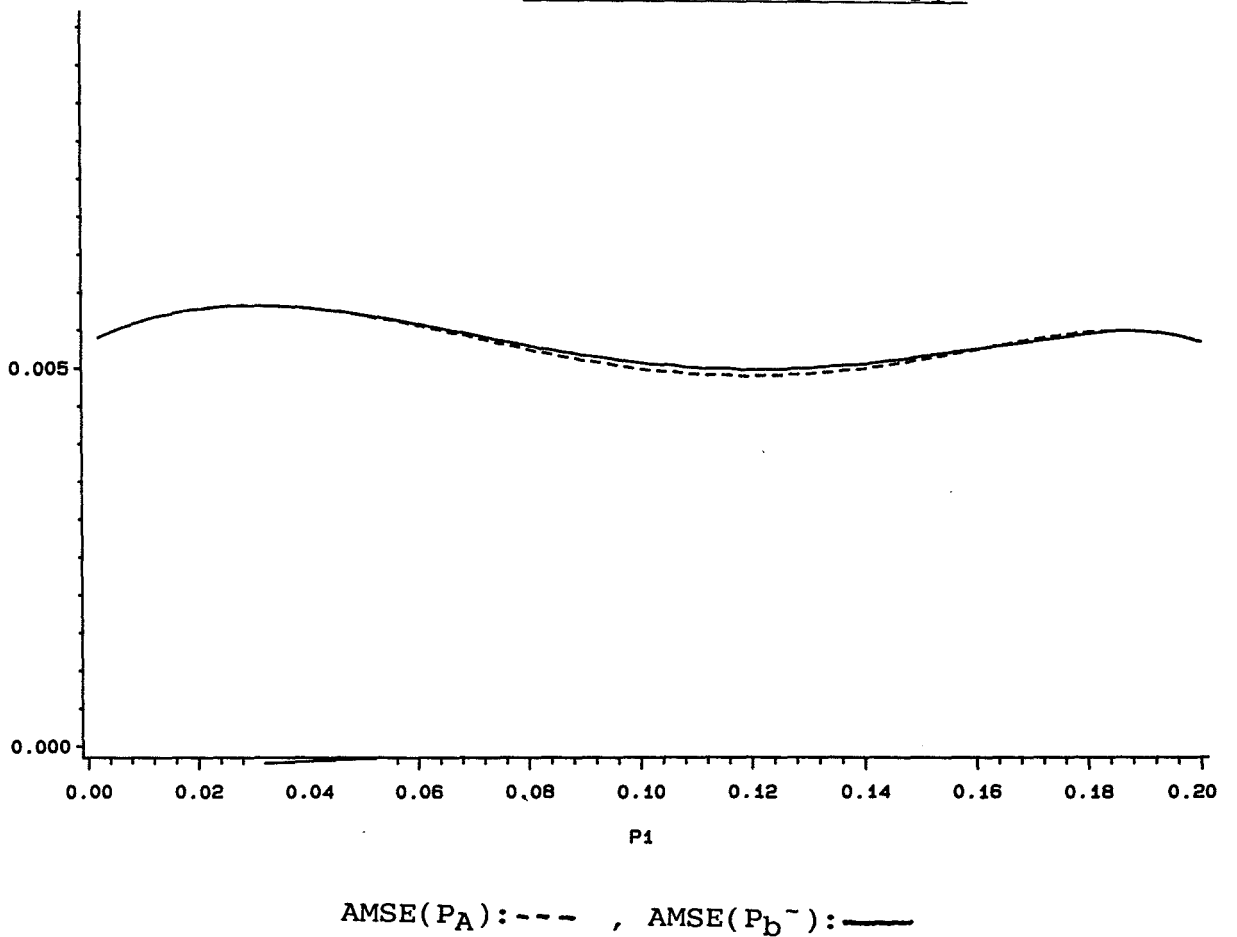
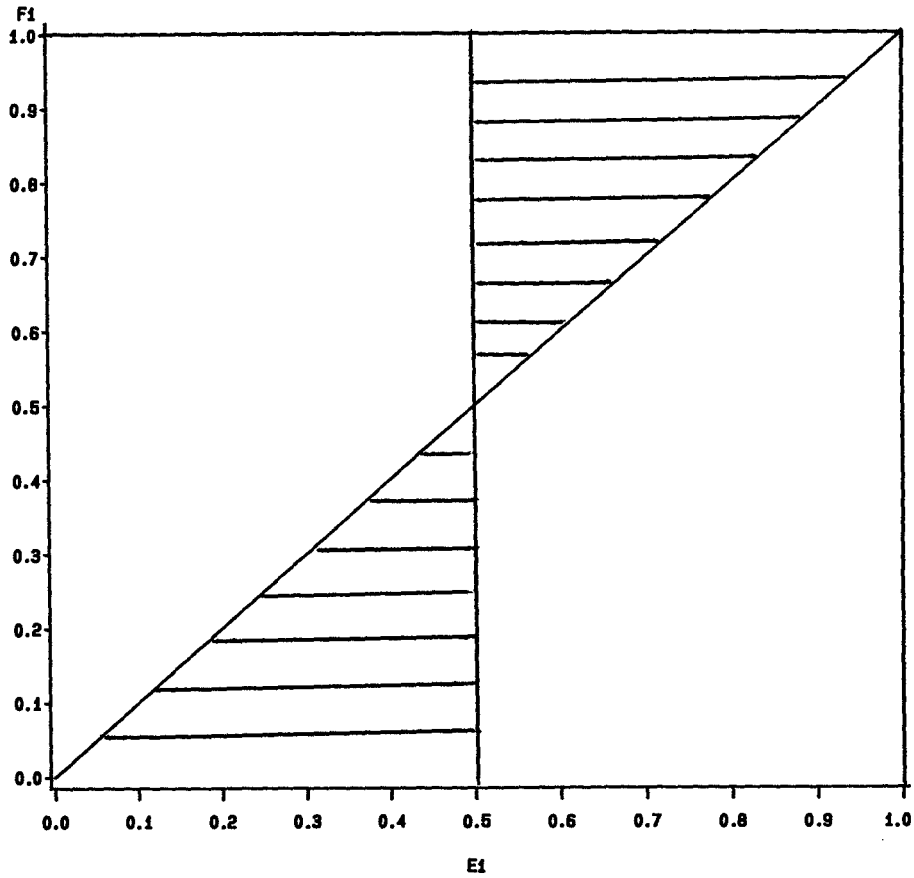
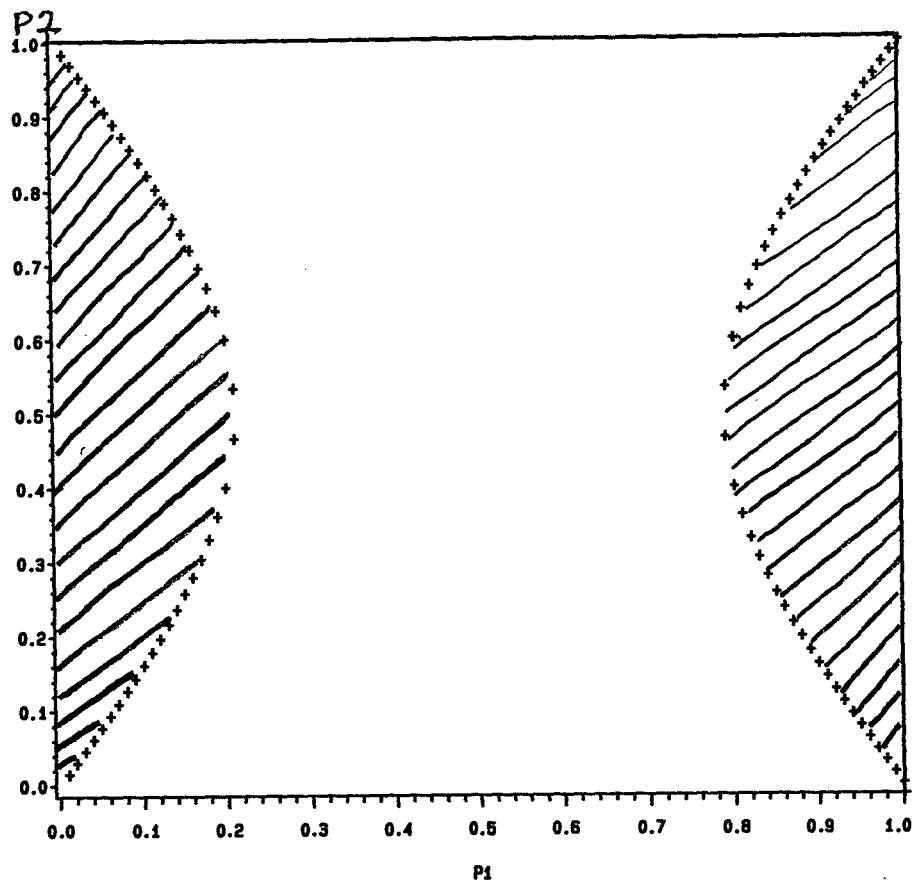



Fig 4.5.1 Effects of different sampling proportions in the old and the new sample for $AMSE(P^*)$.



The shaded area indicates the values of f_1 and e_1 where $AMSE(P^*)$ for $e_1=f_1$ is larger than $AMSE(P^*)$ for $e_1 \neq f_1$.

Fig 4.5.2 Effects of different sampling proportions in the old and the new sample for AMSE(P').



 : values of p_1 and p_2 where AMSE(P') for $f_1=e_1=0.4$ is larger than AMSE(P') for $f_1>e_1=0.4$.


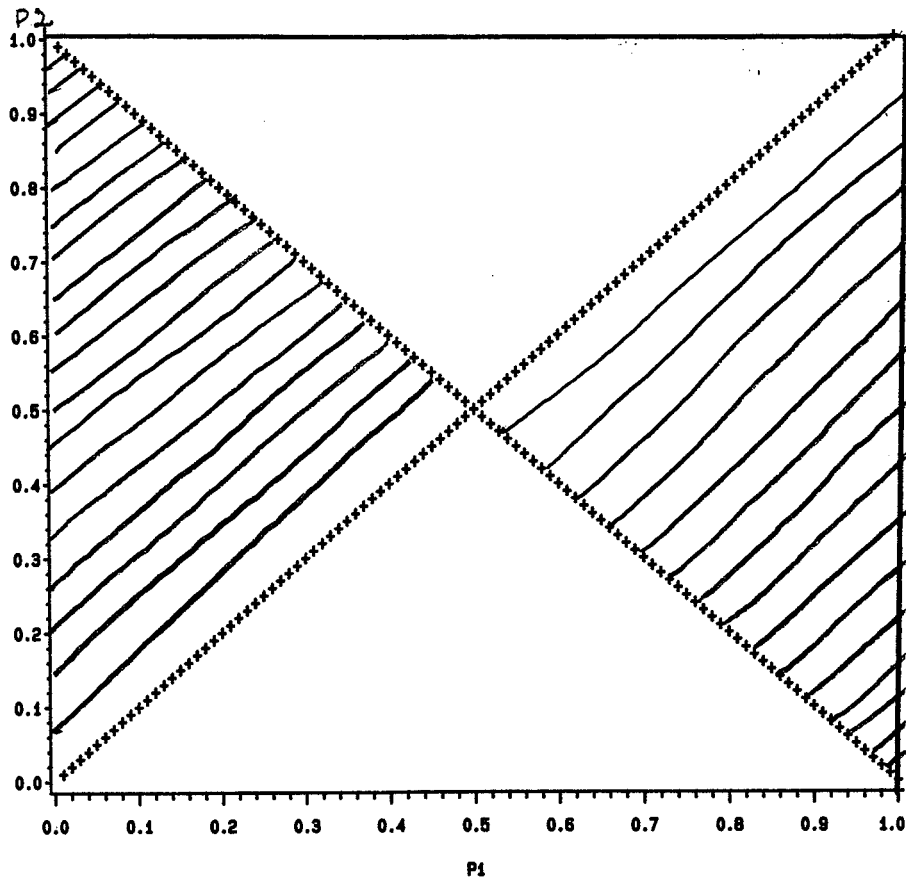

 : values of p_1 and p_2 where AMSE(P') for $f_1=e_1=0.4$ is larger than AMSE(P') for $f_1<e_1=0.4$.

Fig 4.5.3 Effects of different sampling proportions in the old and the new sample for AMSE(P').



 : values of p_1 and p_2 where AMSE(P') for $f_1=e_1=0.5$ is larger than AMSE(P') for $f_1>e_1=0.5$.


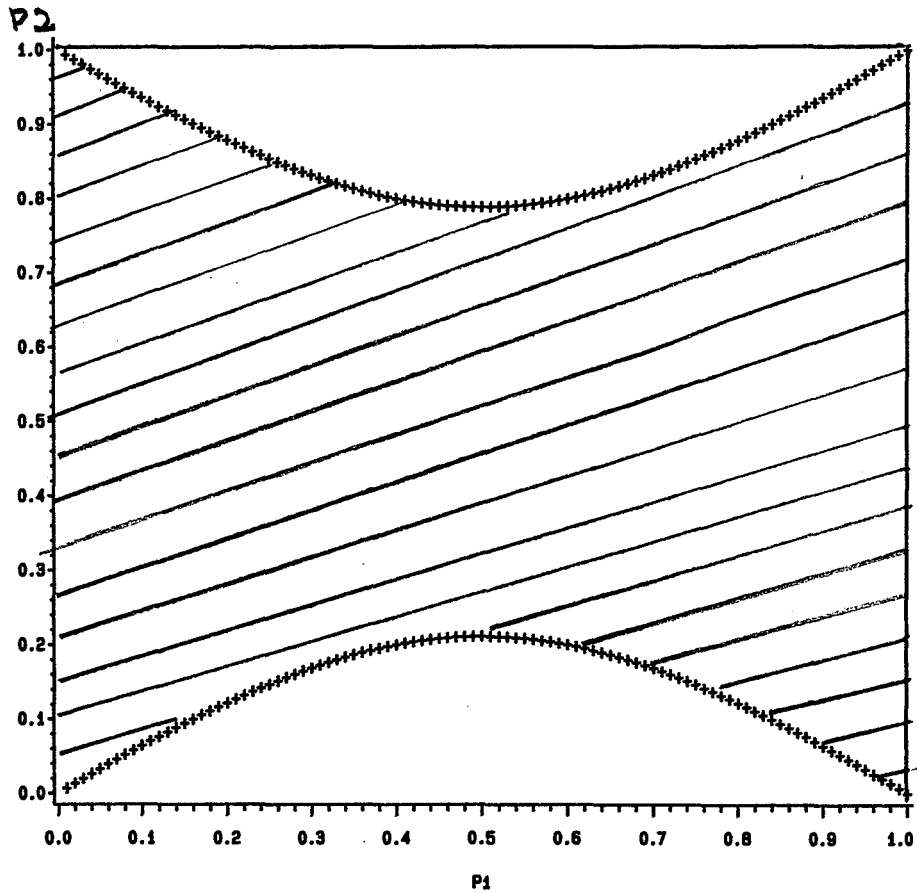


 : values of p_1 and p_2 where AMSE(P') for $f_1=e_1=0.5$ is larger than AMSE(P') for $f_1<e_1=0.5$.

Fig 4.5.4 Effects of different sampling proportions in the old and the new sample for AMSE(P').



 : values of p_1 and p_2 where $AMSE(P')$ for $f_1=e_1=0.6$ is larger than $AMSE(P')$ for $f_1 > e_1 = 0.6$.

 : values of p_1 and p_2 where $AMSE(P')$ for $f_1=e_1=0.6$ is larger than $AMSE(P')$ for $f_1 < e_1 = 0.6$.

5 Concluding remarks.

When making hypothesis testing, the rejection of a true null-hypothesis is usually considered to be a grave error. This consideration calls for adopting a small significance level, such as 0.05 or 0.01. When choosing between two prediction rules, such as P^* and P' , we don't have such prior considerations. This paper has shown that if we adopt a low-level significance test for choosing between P^* and P' this procedure has a high AMSE-risk for large areas of the parameter space.

APPENDIX: Collection of some useful results.

A1. $MSE(p_j^*)$ and $MSE(p_j')$ for the multinomial case.

We begin by evaluating $MSE(p_j^*)$, for the multinomial case. We have that

$$\begin{aligned} MSE(p_j^*) &= E(x_{1.}/m - p_j)^2 = V(x_{1.}/m) + (E(x_{1.}/m) - p_j)^2 \\ &= p_{1.} \cdot (1-p_{1.})/m + (p_{1.} - p_j)^2 \end{aligned}$$

Here we have used the fact that $x_{1.}$ has a binomial distribution with parameters $p_{1.}$ and n .

We proceed by considering $MSE(p_j')$

$$\begin{aligned} MSE(p_j') &= E(x_{1j}/x_{.j} - p_j)^2 = V(x_{1j}/x_{.j}) + \\ &\quad + (E(x_{1j}/x_{.j}) - p_j)^2 \end{aligned}$$

Now, for the expectation of p_j' we have

$$E(p_j') = EE(x_{1j}/x_{.j} \mid x_{.j}) = E(p_j) = p_j$$

By adopting the multivariate deltha method we will evaluate an approximation to the variance of p_j' . We first formulate the method in general terms. Let $\theta = (\theta_1, \dots, \theta_t)'$ be a vector of parameters and let $\theta_n' = (\theta_{n1}', \dots, \theta_{nt}')'$ be a vector of random variables with the same dimension. Assume that θ_n' has an asymptotic normal distribution in the sense that

$$L(n \cdot (\theta_n' - \theta)) \rightarrow N(0, \Sigma(\theta))$$

where L stands for convergence in distribution and $\Sigma(\theta)$ is the asymptotic covariance matrix of θ_n' . Further, let f be a function which has the following expansion as $x \rightarrow \theta$

$$f(x) = f(\theta) + (x - \theta)D_\theta' + o(\|x - \theta\|)$$

where D_θ is the vector of partial derivatives of f evaluated at $x = \theta$. Within this framework, the asymptotic distribution of $f(\theta_n')$ is given by

$$L(n \cdot (f(\theta_n') - f(\theta))) \rightarrow N(0, D_\theta \Sigma(\theta) D_\theta')$$

We proceed by determining the asymptotic variance of p_1' for a 2×2 -table, the argument being the same for p_2' . Put

$$\theta = (n \cdot p_{11}, n \cdot p_{21}, n \cdot p_{12}, n \cdot p_{22})'$$

$$\theta_n' = (x_{11}, x_{21}, x_{12}, x_{22})'$$

$$f(\theta_n') = p_1' = x_{11}/x_{.1} \quad f(\theta) = p_{11}/p_{.1} = p_1$$

It is well-known that the x_{ij} :s have an asymptotic normal distribution and the asymptotic covariance matrix is given by

$$\Sigma(\theta) = \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{21} & -p_{11}p_{12} & -p_{11}p_{22} \\ & p_{21}(1-p_{21}) & -p_{21}p_{12} & -p_{21}p_{22} \\ & & p_{12}(1-p_{12}) & -p_{12}p_{22} \\ & & & p_{22}(1-p_{22}) \end{pmatrix}$$

Computing the elements of D_{θ}

$$\delta f / \delta \theta_{n1}' = \frac{p_{21}}{m \cdot p_{.1}^2}$$

$$\delta f / \delta \theta_{n3}' = - \frac{p_{11}}{m \cdot p_{.1}^2}$$

As $f(\theta_n')$ does not include θ_{n2}' and θ_{n4}' the dimension of D_{θ} is 2×1 . This implies that the asymptotic variance of p_1' is given by

$$m \cdot D_{\theta} \begin{pmatrix} p_{11} \cdot (1 - p_{11}) & -p_{11} \cdot p_{21} \\ -p_{11} \cdot p_{21} & p_{21} \cdot (1 - p_{21}) \end{pmatrix} D_{\theta}'$$

Performing this computation we arrive at

$$(1/n) \cdot (p_{11} \cdot p_{21} / p_{.1}^2) = p_1 \cdot (1 - p_1) / n \cdot p_{.1}$$

A2. Maximum-likelihood estimators for the multinomial case.

Here we derive the M.L-estimators of the probabilities p_j , p_1 , and $p_{.j}$, $j=1,2,\dots,k$, for the multinomial case. The likelihood function for the $2 \times k$ random variables is

$$L = \frac{m!}{\prod \pi x_{ij}!} \prod \pi p_{ij}^{x_{ij}}$$

The essential part of the log-likelihood function is

$$l = \sum_i \sum_j x_{ij} \cdot \log(p_{ij}) \quad i = 0,1 \quad j = 1,2,\dots,k$$

Writing out the summation over the index i we get

$$l = \sum_j (x_{0j} \cdot \log(p_{0j}) + x_{1j} \cdot \log(p_{1j})) =$$

$$\sum_j (x_{0j} \cdot \log(p_{.j} - p_{1j}) + x_{1j} \cdot \log(p_{1j}))$$

Making the substitution $p_{1j} = p_j \cdot p_{.j}$

$$l = \sum_j (x_{0j} \cdot \log(p_{.j} - p_{.j} \cdot p_j) + x_{1j} \cdot \log(p_{.j} \cdot p_j))$$

Maximizing this with respect to p_j by taking derivative and putting this equal to zero

$$\delta l / \delta p_j = x_{1j} \cdot p_{.j} / p_j \cdot p_{.j} - x_{0j} \cdot p_{.j} / (p_{.j} - p_{.j} \cdot p_j) = 0$$

\Leftrightarrow

$$p_j = x_{1j} / x_{.j}$$

For the probabilities $p_{.j}$, we observe that under this sampling plan the vector $(x_{.1}, x_{.2}, \dots, x_{.k})'$ has a multinomial distribution with parameters $m = x_{..}$ and $p_{.1}, p_{.2}, \dots, p_{.k}$. We therefore conclude that the M.L-estimator of $p_{.j}$ is $x_{.j} / m$.

By a similar reasoning we also conclude that the M.L-estimator of $p_{1.}$ is given by $x_{1.} / m$.

REFERENCES

Aitkin , M. (1978): A Simultaneous Test Procedure for Contingency Table Models, J. Roy. Statist. Soc. (C) 141, 195-223

Atkinson, A.C. (1980): A Note on the Generalized Information Criterion for Choice of a Model, Biometrika, 67, 413-418

Benedetti, J.K and Brown, M.D. (1976): Alternate Methods of Building Log-Linear Models, Proceedings of the 9th International Biometric Conference, The Biometric Society, 209-227

Efron, B (1978): Regression and ANOVA with Zero-One Data, JASA, 73, 113-121

Efron, B. (1986): How Biased is the Apparent Error Rate of a Prediction Rule? JASA, 81, 461-470

Fowles, e.b., Freeny, A.E., and Landwehr, J.M. (1988): Evaluating Logistic Models for Large Contingency Tables, JASA, 83, 611-622

Goodman, L.A. (1964): Simultaneous Confidence Limits for Cross-Product Ratios in Contingency Tables, J. Roy. Statist. Soc. (B), 26, 86-102

Hildebrand, D.K., Laing, I.D. and Rosenthal, H. (1977): Prediction Analysis of Cross Classifications, John Wiley & Sons, New York

van Houwelingen, J.C. and Le Cessie, S. (1989): Predictive Value of Statistical Models, Proceedings of the Comp. Stat. Conf. Copenhagen, 1988

Linhart, H. and Zucchini, W. (1986): Model Selection, John Wiley & Sons, New York

1990:1	Holm, S.	Abstract bootstrap confidence intervals in linear models.
1990:2	Holm, S. & Dahlbom, U	On tests of equivalence
1991:1	Olofsson, Jonny	On some prediction methods for categorical data