AN ABSTRACT OF THE THESIS OF

MICHEL GEORGES LEJEUNE for the degree DOCTOR OF PHILOSOPHY
(Date) (Degree)

in STATISTICS presented on July 1, 1974
(Major Department) (Date)

Title: A MAXIMUM LIKELIHOOD APPROACH TO PREDICTION WITH APPLICATIONS TO BINOMIAL AND POISSON POPULATIONS.

Abstract approved: Redacted for privacy
(G. David Faulkenberry)

Given a sample from a population whose distribution belongs to a parametric family we would like to make predictions on the outcome of a statistic (generally the sample sum) of a future sample from the same population. These predictions can be in the form of intervals with some associated level of confidence or in the form of predictive distributions.

A "Maximum Likelihood Predictive Distribution" is introduced which is easily accessible in most regular cases and under conditions similar to those for the consistency of the maximum likelihood estimator converges almost surely to the true distribution of the predicted variable when the observed sample size becomes large.

The new approach is compared to frequentist and Bayesian approaches and to another likelihood approach introduced by R. A. Fisher. The comparisons are conducted for simple random sampling from Poisson populations and from binomial populations.

The various approaches yield quite similar results for all sample sizes and tend to be equivalent to the method using normal approximations when both the observed and future sample sizes tend to infinity such that their ratio remains constant.

It is shown that the Maximum Likelihood Predictive Distribution is almost identical to the Bayesian predictive distribution under prior $\sqrt{\lambda}$ in the Poisson case and prior $\sqrt{p(1-p)}$ in the binomial case.

These approaches are also considered for Poisson and binomial stratified random sampling and the results compared.

A Maximum Likelihood Approach to Prediction With Applications to Binomial and Poisson Populations

by

Michel Georges Lejeune

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed July 1974

Commencement June 1975

APPROVED:

Redacted for privacy

Associate Professor of Statistics

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented July 1, 1974

Typed by Robin Przybilla for Michel Georges Lejeune

ACKNOWLEDGMENT

The author expresses his indebtedness to Dr. David Faulkenberry who initially suggested the problem and whose guidance has been very precious all throughout this research.

To all members of the Department of Statistics the author is grateful, especially to Dr. Justus Seely for his availability and his kindness and to Dr. Donald Pierce for his helpful criticism.

The author also acknowledges the financial assistance of the Department of Statistics under the chairmanship of Dr. Lyle D. Calvin and the Computer Center. Without their financial support this thesis could not have been completed; the Fulbright travel grant was also an important factor in ensuring the possibility of my studies.

Finally, I would like to thank my parents and my friends for their moral support.

TABLE OF CONTENTS

A MAXIMUM LIKELIHOOD APPROACH TO PREDICTION WITH APPLICATIONS TO BINOMIAL AND POISSON POPULATIONS

I. INTRODUCTION

The information contained in a sample drawn from a population whose distribution is known up to some parameter value is used in classical statistical theories to make inferences about this unknown parameter. Prediction theory, however, is concerned about inferences on a future sample to be drawn from the same population, where two sources of uncertainty are combined; one is the uncertainty about the true value of the parameter, the other is due to the randomness of the variables to be "predicted".

A prediction statement well known to statisticians is found in regression theory where one seeks an interval that will contain an outcome of the dependent variable at a given value of the independent variable with some given probability. Another familiar problem of the predictive type is the so-called "rule of succession" which Laplace was concerned with. (see Fisher (1959).).

In this thesis we restrict our interest to inferences on a one-dimensional statistic of a future sample, typically the sample sum. The basic type of prediction that will be considered is in the form of a "prediction interval" defined as follows:

<u>Definition</u>: Let $\underset{\sim}{X} = (x_1, x_2, \ldots, x_n)$ be a random sample from a population with distribution $F_X(x;\theta)$ where $\theta \in \Theta$ is unknown. Denote $\underset{\sim}{Y} = (y_1, y_2, \ldots, y_m)$ m "future" outcomes from the same population and

let $Z = h(\underset{\sim}{Y})$ be a one-dimensional function of these outcomes. Suppose we find two functions $L_\alpha(\underset{\sim}{X})$ and $U_\alpha(\underset{\sim}{X})$ such that:

$$Pr_\theta\{L_\alpha(\underset{\sim}{X}) < Z < U_\alpha(\underset{\sim}{X})\} = \alpha \tag{1.1}$$

for all $\theta \varepsilon \Theta$ and with α independent of θ. Then the intervals $[L(\underset{\sim}{X}), U(\underset{\sim}{X})]$ are said to be a family of prediction intervals on Z given $\underset{\sim}{X}$ of confidence level α.

For an outcome $\underset{\sim}{x}$ of $\underset{\sim}{X}$ the values $L(\underset{\sim}{x})$ and $U(\underset{\sim}{x})$ are respectively, the α-lower and upper prediction limits of Z given $\underset{\sim}{x}$.

A favorable situation occurs when there exists a function $f(\underset{\sim}{X}, Z)$ whose distribution is independent of θ and which is monotone and continuous in z. A case of this type is the prediciton of the mean $\bar{Y}$ of a future sample of size m from a Normal population. Since

$$\frac{\bar{Y} - \bar{X}}{s_X\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n - 1) \text{ , where } s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}$$

a prediction interval follows immediately. Moreover for an observed $\underset{\sim}{X} = \underset{\sim}{x}$ the determination of a prediction interval at any confidence level reduces to the computation of fractiles for the random variable with distribution

$$\bar{x} + t(n - 1)\ s_x \sqrt{\frac{1}{n} + \frac{1}{m}} \quad .$$

Consequently we call this distribution a "predictive distribution" of $\tilde{Y}$ given $\underset{\sim}{x}$.

As another example, Fisher (1959), who devoted much attention to the prediction problem, shows that the ratio of the sample sums X and Y is independent of θ when sampling from an exponential distribution.

A more general type of situation is when the distribution function of $\underset{\sim}{X}$, $F_{\underset{\sim}{X}}(\underset{\sim}{x};\theta)$ yields a pivotal quantity. In this case the fiducial argument as introduced by Fisher (1959) may be applied. According to this argument the "logical status" of the parameter is changed from "one in which nothing is known to the status of a random variable having a well-defined distribution". Let $G(\theta|\underset{\sim}{x})$ be this distribution, then a predictive distribution on Y is defined as

$$F_{\underset{\sim}{Y}}(\underset{\sim}{y};x) = \int_{\Theta} F_{\underset{\sim}{Y}|\theta}(\underset{\sim}{y};\theta)dG(\theta|\underset{\sim}{x}) \quad .$$

A quite illustrative example is given by Kalbfleisch and Sprott (1969) for a life testing problem.

In many situations however there is no fiducial argument available and some alternative general method must be used.

A first class of methods are the "frequentist methods", i.e., methods where the probability statement 1.1 of the definition is derived in classical ways from the probability distributions. Faulkenberry (1972) gives a "frequentist conditional" approach which is fairly general. Nelson (1970) has shown a hypothesis testing approach in some special cases. The Bayesian approach is certainly the most general in the sense that it can be used in most situations; Aitchison and Sculthorpe (1965) give the Bayesian formulation as well as a decision theoretic formulation for the prediction problem. A quite different class of approaches is based on likelihood statements as was first suggested by Fisher (1959) and further emphasized by Sprott and Kalbfleisch (1969,1971). It is to be noted that the concept of likelihood is controversial (see Kempthorne (1969) and Barnard (1969).).

On the basis of its asymptotic properties a "Maximum Likelihood Predictive Distribution" (MLPD) which is available and easy to derive in regular cases is proposed in this thesis. Comparisons of the frequentist conditional, Bayesian and likelihood approaches are conducted in the special cases of binomial and Poisson populations for which special problems arise from the discreteness. In both cases it is seen that the MLPD is very much in agreement with predictions available from other approaches.

II. PREDICTION METHODS

In this chapter the frequentist conditional and the Bayesian approaches are presented. Some results specific to discrete random variables are derived for the frequentist conditional method. The emphasis is put on likelihood approaches. A close examination of the likelihood proposed by Fisher leads to the definition of another type of likelihood called the "Prediction Likelihood Function". It is shown that under certain regularity conditions this PLF converges in probability to the true density of the predicted variables up to a proportionality constant when the observed sample size tends to infinity. This asymptotic property leads to the introduction of the "Maximum Likelihood Predictive Distribution" (MLPD) whose density is defined to be proportional to the PLF whenever the latter is integrable.

II.1 Frequentist conditional approach

II.1.1 General formulation

Let us formulate the solution given by Faulkenberry (1973) in a slightly different way.

Let $\underset{\sim}{X}$ be a random vector with distribution $F_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)$ and Z a (one-dimensional) random variable independent of $\underset{\sim}{X}$ with distribution $F_Z(z|\theta)$, where θ is the same for both distributions. Suppose $\underset{\sim}{T}$ is a sufficient statistic for the joint distribution of $(\underset{\sim}{X},Z)$ and there exists a region $R'(\underset{\sim}{t})$ in R^1 such that

$$\int_{z \in R'(\underset{\sim}{t})} dF_{Z|\underset{\sim}{T}}(z|\underset{\sim}{t}) = \alpha .$$

Suppose further that there exists an interval $R(\underset{\sim}{x})$ in R^1 such that

$$z \in R(\underset{\sim}{x}) \iff z \in R'(\underset{\sim}{t}) .$$

Then

$$\Pr_\theta\{Z \in R(\underset{\sim}{X})\} = \alpha \quad \text{for all} \quad \theta \in \Theta .$$

That is, for any outcome $\underset{\sim}{x}$ of $\underset{\sim}{X}, R(\underset{\sim}{x})$ is an α-confidence prediction interval.

Note that the only theoretical restriction is the existence of $R(\underset{\sim}{x})$ as an interval on R^1.

Olsen (1974) gave various conditions under which this method has an easy solution, especially for problems where the sample sums X and Z are sufficient statistics for the observed and the future sample respectively, and where T is chosen to be $X + Z$. Olsen defines further a predictive distribution for Z, but as will be seen in the next section some problems arise when dealing with discrete distributions.

Note: One may also think of conditioning the observed variable $\underset{\sim}{X}$ (or a function of $\underset{\sim}{X}$) on a sufficient statistic $\underset{\sim}{T}(\underset{\sim}{X},Z)$. $\underset{\sim}{T}$ can now be seen as a parameter and be given a confidence interval in the classical manner, hoping that this interval can be translated into an interval on Z. In the case of sample sums the solution for $Z|X + Z$ is equivalent to the solution for $X|X + Z$.

II.1.2 Discrete case

In the discrete case we face the same difficulty in prediction as for confidence limits for the parameter of a discrete distribution (see Stevens (1950) and Pratt (1965).). This has been described for the binomial prediction by Thatcher (1964) and in this section we will generalize some of his results.

Let us first give a definition that will be convenient when stating results concerning discrete distributions.

Definition 2.1: The "β-upper fractile" of a distribution is the smallest real number u such that for a random variable X having this distribution

$$\Pr\{X \leq u\} \geq \beta.$$

The "β-lower fractile" is the largest real number ℓ such that

$$\Pr\{X \geq \ell\} \geq \beta.$$

To simplify the following developments assume that the discrete random variables X,Y take values on $I_x = \{0,1,2,\ldots,n_x\}$ and $I_y = \{0,1,2,\ldots,n_y\}$ respectively where n_x and n_y can be infinite. The sample sums are X and Y, and $T = X + Y$ is sufficient; we are looking first for an upper prediciton limit on Y.

Once a confidence level α has been chosen, the determination of an upper limit in the conditional problem $Y|T$ cannot be accomplished at the exact confidence level. In fact for $T = t$ we can only choose an integer-valued function $h(t)$ such that

$$\Pr\{Y \leq h_\alpha(t) | T = t\} \geq \alpha \tag{2.1}$$

and

$$\Pr\{Y \geq h_\alpha(t) \mid T = t\} > 1 - \alpha \quad . \tag{2.2}$$

As this is done for all t, the (x,y) plane is partitioned into two sets of points R and $\bar{R}$, where

$$R = \{(x,y): \; y \leq h_\alpha(x + y)\}$$

By taking the expectations over the density of T of both sides of (2.1) and (2.2) for some θ we find that for the unconditional problem we have

$$\Pr_\theta\{Y \leq h_\alpha(T)\} \geq \alpha \quad \text{and} \quad \Pr_\theta\{Y \geq h_\alpha(T)\} > 1 - \alpha \quad \text{for all} \quad \theta.$$

The first relationship can be written as

$$\Pr_\theta\{(X,Y) \; \varepsilon \; R\} \geq \alpha \; .$$

Suppose now that it is possible to define R as

$$R = \{(x,y): \; y \leq u(x)\}$$

then $\Pr_\theta\{Y \leq u(X)\} \geq \alpha$ for all θ (2.3)

and thus $u(X)$ defines a family of "conservative" α-upper prediction limits.

We show now that with some rather standard assumptions on $h(t)$ the function $u(x)$ exists and can be determined in a convenient way.

Theorem 2.2: Assume that the functions $h_\alpha(t)$ and $t-h_\alpha(t)$ are non-decreasing for any α. Then the α-upper prediction limit of Y given $\underset{\sim}{x}$ is the unique value U such that

$$\Pr\{Y \leq U | T = U + x + 1\} \geq \alpha$$

$$\text{and} \quad \Pr\{Y \leq U - 1 | T = U + x\} < \alpha \tag{2.4}$$

Proof: From the assumptions we have

$$h_\alpha(t) \leq h_\alpha(t + 1)$$

$$\text{and} \quad t - h_\alpha(t) \leq t + 1 - h_\alpha(t + 1)$$

$$\text{or} \quad h_\alpha(t + 1) \leq h_\alpha(t) + 1$$

Thus $h_\alpha(t)$ is incremented by either 0 or 1. Let us see what this means for the boundary of the region R (see figure 2.1). Since the points $x + y = \text{constant}$ are located on a left diagonal, the previous results mean that, going away from the origin along the boundary of R, each point is followed by a point located either to its right or just above. Obviously for this to happen the assumptions are necessary.

Then for a given x, we look for the highest point in R on the vertical x, i.e., we look for U such that $(x,U) \in R$ and $(x,U+1) \notin R$, or equivalently

$$\Pr\{Y \geq U | U + x\} > 1 - \alpha$$

and $$\Pr\{Y \geq U + 1 | U + x + 1\} \leq 1 - \alpha$$

which can be readily written in the form (2.4).

Corollary 2.3: Suppose that for all $k \in K \subset I$ there exists a discrete distribution P_k over I_y such that for a r.v. Z having this distribution:

$$\Pr\{Z \leq y;k\} = \Pr\{Y \leq y | T = y + k\} .$$

Then if $x + 1 \in K$ the (conservative) α-upper prediction limit is the α-upper fractile of the distribution P_{x+1}.

This result is immediate since (2.4) is equivalent to

$$\Pr\{Z \leq U;\ x + 1\} \geq \alpha$$

$$\Pr\{Z \leq U - 1;\ x + 1\} < \alpha .$$

Note that $F_k(y) = \Pr\{Y \leq y | Y + k\}$ is always non-decreasing in y. Suppose that $\Pr\{Y \leq y + 1 | y + 1 + k\} < \Pr\{Y \leq y | y + k\} = \alpha$, then for an α-upper prediction the points (k,y) and $(k,y+1)$ would be respectively a boundary and an interior point of R, which contradicts the assumptions in theorem 2.1. Moreover for the finite discrete case $F_k(y)$ attains the value 1 for all $k \in I_x$, namely when $y = n_y$.

For the infinite case it will be generally true that $F_k(y) \to 1$ when $y \to \infty$ (or equivalently $\Pr\{X \leq x|T\} \to 0$, when $T \to \infty$). So that in most situations the distribution P_x exists for all $x \in I_x$.

For an α-lower prediction limit we define the set of points R' in the (x,y)-plane $R' = \{(x,y): y \geq h^1_\alpha(x + Y)\}$, where the function h^1 is such that

$$\Pr\{Y \geq h^1(t)|T = t\} \geq \alpha$$

and

$$\Pr\{Y \leq h^1(t)|T = t\} > 1 - \alpha.$$

We establish now the result for a lower prediction limit corresponding to theorem 2.2.

<u>Theorem 2.4</u>: Assume that $h^1_\alpha(t)$ and $t - h^1_\alpha(t)$ are non-decreasing for all α. Then the α-lower prediction limit of Y given x is the unique value L such that

$$\begin{aligned} &\Pr\{Y \leq L|T = L + x\} > 1 - \alpha \\ &\Pr\{Y \leq L - 1|T = L + x - 1\} \leq 1 - \alpha \end{aligned} \tag{2.5}$$

<u>Proof:</u> The boundary of R' has the same shape as for R, (see figure 2.1) and thus we have to find the lowest value L for which (L,x) is in R' which readily leads to (2.5).

<u>Corollary 2.5</u>: With the same assumptions as in corollary 2.3 L is the α-lower fractile for the distribution P_x.

In conclusion we see that we had to define limits in a conservative way. For example if we were to choose for upper limit the lowest points vertically on the upper boundary of R then statement (2.3) would no longer be true. It is to be noted that what is described as a conservative procedure for the upper limit is an anti-conservative procedure when applied to the lower limit, and vice-versa.

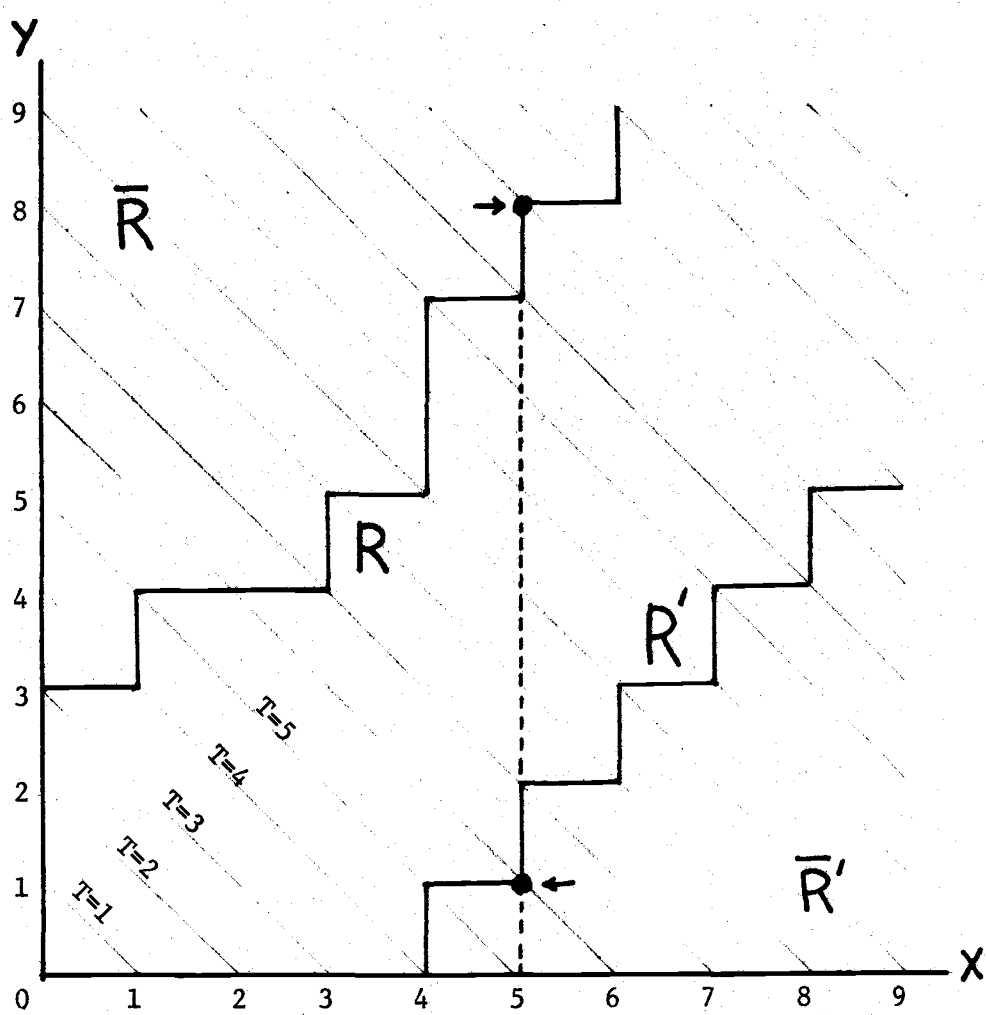


FIGURE 2.1

II.2 Bayesian approach

Suppose a prior $\pi(\theta)$ is given for the parameter $\theta \varepsilon \Theta \subset R^k$ which has here the status of a random variable. Then x being an observation of a random variable (or vector) X, we define the posterior density of θ given x as

$$\pi(\theta|x) = \frac{f_X(x|\theta)\pi(\theta)}{\int_\Theta f_X(x|\theta)\pi(\theta)d\theta}$$

whenever $\int_\Theta f_X(x|\theta)\pi(\theta)d\theta$ exists.

The Bayesian predictive density of Z given x is

$$h_Z(z|x) = \int_\Theta f_Z(z|\theta)\pi(\theta|x)d\theta$$

$$= \frac{\int_\Theta f_Z(z|\theta)f_X(x|\theta)\pi(\theta)d\theta}{\int_\Theta f_X(x|\theta)\pi(\theta)d\theta} .$$

This density $h_Z(z|x)$ can be viewed as the density of the conditional distribution $Z|X$, where the numerator is the joint distribution of Z and X, and the denominator is the marginal distribution of X.

The α-upper (resp. lower) fractile of this distribution is called the "α-Bayes upper (resp. lower) prediction limit for Z given $X = x$ under prior π ".

<u>Case of uniform prior</u>: Assuming that the integral $\int_\Theta f_X(x|\theta)d\theta$ exists, one can choose $\pi(\theta) = 1$ for all $\theta \varepsilon \Theta$, and

$$h_Z(z|x) \propto \int_{\Theta} f_Z(z|\theta) f_X(x|\theta) d\theta \ .$$

One should note that to integrate out the unknown parameter θ in this way corresponds to mixing the density of Z, $f_Z(z|\theta)$, with a density for θ proportional to its likelihood function inferred from the observation.

II.3 Likelihood approaches

II.3.1 Fisher's Likelihood

This method was first introduced by Fisher (1959) on a 2×2 contingency table and extended by Kalbfleisch and Sprott (1969) to the general prediction problem.

Definition 2.6: Suppose X and Z are independent random vectors with densities $f_X(x;\theta)$ and $f_Z(z;\theta)$ respectively where $\theta \in \Theta$ is known to have the same value for both densities (By "density" it is meant a probability density function of either the continuous or the discrete type). We define $R_X(\theta;x)$ as the relative likelihood of θ based on the outcome x of X, where

$$R_X(\theta;x) = f_X(x;\theta)/\sup_{\theta \in \Theta} f_X(x;\theta) ,$$

and in the same way we define $R_Z(\theta;z)$.

If X is observed and Z is to be predicted, then "Fisher's likelihood" of Z given $X = x$ is defined by

$$\tilde{L}(z;x) = \sup_{\theta \in \Theta} R_X(\theta;x) \ R_Z(\theta;z) \quad .$$

Fisher's reasoning for the use of this measure is that the likelihood function of the "aggregate" of two independent sets of data is the product of the likelihood of the two sets; here the unknown parameter θ is taken to be the "most plausible value" for each given conjecture (x,z). Fisher emphasizes the symmetry in x and z and the fact that "the same measure may be taken to be the likelihood of the hypothesis" that θ is the same for the distribution of X and Z.

Kalbfleisch and Sprott (1969) in their extension of the method tried to justify Fisher's choice. But only ulteriorally Kalbfleisch (1971) gave a theoretical foundation for this likelihood. Doing this he brings up the fact that it is essentially a measure of the plausibility of the hypothesis that X and Z come from members of their respective parametric families of distributions with the same value for θ.

Then following an argument used by Nelson (1970) in his hypothesis testing approach, Kalbfleisch states that since θ is known to be the same the likelihood has to measure the plausibility of a value z of Z. The latter statement is not entirely convincing and rather vague. Nevertheless this likelihood would certainly be appropriate to test whether a given value of Z, were it observed subsequently to an outcome x of X, would support the hypothesis that $\theta_X = \theta_Z$. In this case the symmetry in x and z (say when X and Z are identical statistics) is necessary. But for our prediction problem the difference in status of X and Z, known and unknown, should induce an asymmetry in the likelihood of Z since such a measure is relative and compares all possible values of Z, with x being fixed.

Let us now turn to some properties of Fisher's Likelihood. In the following we assume that there exists $\hat{\theta}_{XZ}(x,z)$ or briefly $\hat{\theta}_{XZ}$ such that

$$R_X(\hat{\theta}_{XZ};x)R_Z(\hat{\theta}_{XZ};z) = \sup_{\theta \in \Theta} R_X(\theta;x)R_Z(\theta;z)$$

<u>Property 1:</u> $\hat{\theta}_{XZ}(x,z)$ as defined above is the maximum likelihood estimator based on X and Z jointly, since

$$R_X(\theta;x)R_Z(\theta;z) = k(x)h(z)\ f_X(x;\theta)\ f_Z(z;\theta) \quad .$$

<u>Property 2:</u> Assume further that for almost every x [1] there exists a MLE for θ denoted $\hat{\theta}_X(x)$, i.e. $\hat{\theta}_X(x)$ satisfies

$$R_X(\hat{\theta}_X(x);x) = \sup_{\theta \in \Theta} R_X(\theta;x) = 1 \; .$$

Assume the same for Z and denote the MLE of θ based on Z by $\hat{\theta}_Z(z)$. Let $\hat{\Theta}_X$ be the set image of the function $\hat{\theta}_X(x)$ for the set of x's such that $f_X(x;\theta_o) > 0$ for some θ_o .

Then a necessary and sufficient condition for $\tilde{L}(z;x)$ to attain for a.e. x a maximum (equal to 1) for a value of z in the sample space $\mathcal{Z}$ of Z is that for all $\theta \in \hat{\Theta}_X$ there exists a $z \in \mathcal{Z}$ such that

$$\theta = \hat{\theta}_Z(z) \qquad \text{or} \qquad z = \hat{\theta}_Z^{-1}[\theta] \; .$$

Then the value of z yielding the maximum is $\hat{z} = \hat{\theta}_Z^{-1}[\theta(x)]$.

<u>Proof:</u> (necessary) Since $R_X(\theta;x)$ and $R_Z(\theta;x)$ are at most equal to 1 for all x,y,θ the only way to get $\tilde{L}(z;x) = 1$ is to have both above terms equal to 1. Because x is fixed we have to take $\hat{\theta}_{XZ} = \hat{\theta}_X(x)$ in order to get $R_X(\hat{\theta}_{XZ};x) = 1$ and then we must find z such that $R_Z(\hat{\theta}_X(x);z) = 1$, i.e. there must exists a $\hat{z}$ such that

[1](A condition on a family of distributions $\{P_\theta(x)\ ,\ \theta \in \Theta\}$ is said to hold "for almost every x" when it holds for all x except for a set A such that $P_\theta(A) = 0$ for all $\theta \in \Theta$).

$$\hat{\theta}_Z(\hat{z}) = \hat{\theta}_X(x) \tag{2.6}$$

and this for a.e. x . It is obvious that the condition is sufficient.

This property makes it easy to find the maximum $\hat{z}$. In fact even when $\hat{z}$ is not in the sample space $\mathcal{Z}$ it can be obtained by solving (2.6) . For example for a Poisson distribution with unknown parameter λ and $x = \sum_{i=1}^{n} x_i$ and $z = \sum_{i=1}^{m} z_i$, z is such that $\hat{\lambda}(x) = \hat{\lambda}(z)$,

i.e. $\quad x/n = \hat{z}/m \quad$ or $\quad \hat{z} = mx/n$.

In general $\hat{z}$ will be fractional.

<u>Property 3:</u> Asymptotic behavior when the observed sample size $n \to \infty$

Let us first show what happens when the observed sample size n becomes large and the future sample size m remains finite for the binomial case studied by Fisher (1959).

We denote by x the number of observed successes, y the number of predicted successes and p the unknown proportion of successes. Then

$$R_x(p;x) = \frac{p^x q^{n-x}}{(x/n)^x (1 - x/n)^{n-x}}$$

$$R_y(p;y) = \frac{p^y q^{m-y}}{(y/m)^y (1 - y/m)^{m-y}}$$

and $\tilde{L}(y;x) = R_X(\hat{P}_{XY};x)R_y(\hat{P}_{XY};y)$ with $\hat{P}_{XY} = (x + y)/(m + n)$.

From the consistency property of the MLE we know that $x/n \to p_o$

almost surely, where p_o is the true proportion. We will see in section II.3.2 that $\hat{P}_{XY}$ tends also almost surely to p_o. Thus the maximum $\hat{y} = mx/n$ tends to mp_o, and since we are interested in values of y around the maximum, we look at $\tilde{L}(y;x)$ for y finite.

Writing

$$\hat{P}_{XY} = (x/n)(1 + y/x)/(1 + m/n)$$

and

$$\hat{q}_{XY} = (1 - x/n)[1 + (m - y)/(n - x)]/(1 + m/n)$$

we get

$$R_X(\hat{P}_{XY};x) = \frac{(1 + y/x)^x}{(1 + m/n)^x} \frac{[1 + (m - y)/(n - x)]^{n - x}}{(1 + m/n)^{n - x}}$$

Noting that $n - x \to \infty$ when $n \to \infty$, since $n - x = n(1 - x/n)$, we obtain

$$\lim_{n\to\infty} R_X(P_{XY};x) = \frac{e^y e^{m - y}}{e^m} = 1 ;$$

thus

$$\lim_{n\to\infty} \tilde{L}(y;x) = \frac{p_o^y q_o^{m - y}}{(y/m)^y (1 - y/m)^{m - y}} .$$

This shows the deceiving behavior of this measure, because one would expect an appropriate measure to tend to be proportional to the density of Y, $f_Y(y;p_o)$.

We will see in section II.3.2 that under certain regularity conditions Fisher's likelihood will tend to

$$f_Z(z;\theta_o) / \sup_{\theta \in \Theta} f_Z(z;\theta)$$

However it will be seen for the special cases studied subsequently that when both m and n become large Fisher's Likelihood yields the same kind of inference as the usual normal approximation.

We now introduce another "likelihood" that will have the required asymptotic properties so that more emphasis will be given to it.

II.3.2 The prediction likelihood function

Definition 2.7: Let X be a random vector with density $f_X(x;\theta)$, $\theta \varepsilon \Theta$ and Z a random vector with density $f_Z(z;\theta)$, where θ is known to be the same for both densities. Then we define as the "prediction likelihood function (PLF) of Z having observed X = x" the function

$$\hat{L}(z;x) = k(x) \sup_{\theta \varepsilon \Theta} f_X(x;\theta) f_Z(z;\theta) ,$$

where k(x) is a normalizing constant chosen such that

$$\sup_{z \varepsilon \mathcal{Z}} \hat{L}(z;x) = 1 \quad \text{and} \quad \mathcal{Z} = \{z : f_Z(z;\theta) > 0 \quad \text{for some} \quad \theta\} .$$

Note that the value of θ that yields the supremum is the MLE based on x and z as in Fisher's method.

Interpretation: The likelihood thus defined has a direct interpretation in the discrete case. The likelihood ratio of two values of Z , z_1 and z_2 , is equal to the ratio of the highest probability of occurrence of (x,z_1) and (x,z_2) for all possible states of nature. In other words the prospective values of Z are compared according and proportionally to the highest probability of observing them in combination with the event $X = x$.

For the continuous case "probability" has to be replaced by "probability density".

Assuming that Z is one-dimensional we can give a geometric representation of the PLF. Since x is fixed, $k(x)f_X(x;\theta)f_Z(z;\theta)$ is a parametric family of curves, $P_\theta(z)$, whose envelope is the PLF.

If we were to look at likelihood intervals by cutting the PLF by horizontal lines (see for example Hudson(1971).), we would include the set of values of z such that the combined observation (x,z) has a probability (resp. probability density) larger than a chosen level for at least one θ. Or, equivalently, we exclude those values that give to the event (x,z) a probability below the chosen level whatever the state of nature is.

Example: We want to make predictions from a normal population with unknown mean and unknown variance.

Let $\underset{\sim}{x} = (x_1,x_2,x_3,\ldots,x_n)$ denote the observed sample, X the sample sum and S_x^2 the sum of squares. We want to predict the sum, Y , of a sample of size m . We have

$$f_{\underset{\sim}{X}}(\underset{\sim}{x};\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\} \quad ,$$

$$f_Y(Y;\mu,\sigma^2) = (2\pi m\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2m\sigma^2}(Y - m\mu)^2\} \quad ,$$

and

$$f_{\underset{\sim}{X}}(\underset{\sim}{x};\mu,\sigma^2)f_Y(Y;\mu,\sigma^2) = m^{-\frac{1}{2}}(2\pi\sigma^2)^{-(n+1)/2} \exp\{-\frac{1}{2\sigma^2}[\sum x_i^2 + \frac{Y^2}{m} - 2\mu(X + Y) + (n+m)\mu^2]\} \tag{2.7}$$

Taking the logarithm of this expression and differentiating with respect to μ and σ^2, we obtain

$$\hat{\mu} = (X + Y)/(m + n)$$

and

$$\hat{\sigma}^2 = \frac{1}{n+1}\left[\sum_{i=1}^{n} x_i^2 + \frac{Y^2}{m} - \frac{(X+Y)^2}{m+n}\right]$$

$$= \frac{1}{n+1}\left[S_x^2 + \frac{n}{m(m+n)}\left(Y - \frac{m}{n}X\right)^2\right] .$$

Noting that the exponential term in (2.7) is a constant for $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, the PLF is

$$\hat{L}(Y;x) \propto \left[\frac{n}{m(m+n)}\left(Y - \frac{m}{n}X\right)^2 + S_x^2\right]^{-(n+1)/2}$$

$$\propto \left[\frac{n}{m(m+n)S_x^2}\left(Y - \frac{m}{n}X\right)^2 + 1\right]^{-(n+1)/2}$$

$$\propto \left[\frac{t^2}{n} + 1\right]^{-(n+1)/2} \quad \text{where} \quad t^2 = \frac{n^2}{m(m+n)S_x^2}\left(Y - \frac{m}{n}X\right)^2$$

and defining $s'^2_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then

$$t^2 = \frac{n}{m}\,\frac{(Y - m\bar{x})^2}{(m+n)s'^2_x} \quad \text{or} \quad t = \frac{(\bar{Y} - \bar{x})}{s'_x\sqrt{\frac{1}{n} + \frac{1}{m}}} .$$

Thus the PLF is proportional to a central t-distribution with n degrees of freedom if we rescale Y in

$$\frac{\bar{Y} - \bar{x}}{s'_x\sqrt{\frac{1}{n} + \frac{1}{m}}} .$$

This is the standard result inferred from the fiducial argument except for the degrees of freedom n instead of $n - 1$ and s'_x with a divisor n instead of $n - 1$.

Note that we can't define Fisher's likelihood for Y alone. We cannot disassociate inference on Y and S_Y^2 because of the factor $f_Y(y;\hat{\theta}(y))$ for which $\hat{\theta}$ is a function of both Y and S_Y^2.

When the variance is known (and equal to 1) we have

$$f_{\underset{\sim}{X}}(\underset{\sim}{x};\mu)f_Y(Y;\mu) \propto \exp\{-\tfrac{1}{2}[\frac{(X - n\mu)^2}{n} + \frac{(Y - m\mu)^2}{m}]\}$$

$$\hat{L}(Y;\underset{\sim}{x}) \propto \exp\{-\tfrac{1}{2}[\frac{n}{m(m+n)} (Y - \frac{m}{n} X)^2]\} \text{ , since } \hat{\mu} = (X+Y)/(m+n)$$

$$\hat{L}(Y;\underset{\sim}{x}) \propto \exp\{-\tfrac{1}{2} \frac{(\bar{Y} - \bar{x})^2}{\frac{1}{m} + \frac{1}{n}}\} ,$$

which is equivalent to the frequentist result. For Fisher's method we get the same result since we have to divide by

$$f_Y(Y;\hat{\mu}_Y) \propto \exp\{-\tfrac{1}{2}(Y - m\hat{\mu}_Y)\} = \text{constant} .$$

Properties and theorems for the PLF:

Property 1: The PLF, $\hat{L}(z,x)$, depends on x only through its sufficient statistic with respect to θ. This follows immediately from the definition.

Property 2: The following theorems show that in regular cases the PLF tends to be proportional to the true density function of the random variable to be predicted, when the observed sample size increases.

Theorem 2.8: Let $\{f(x;\theta),\ \theta \in \Theta\}$ and $\{g(y;\theta),\ \theta \in \Theta\}$ be two parametric families of probability density functions with $\Theta \subset R^1$. Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x;\theta_o)$ and assume the following conditions of regularity:

i) $E_{\theta_o}\{|(\partial/\partial\theta) \log f(X,\theta)|\} < \infty \quad \forall\theta \in \Theta$

ii) $E_{\theta_o}\{|(\partial^2/\partial\theta^2) \log f(X,\theta)|\} < \infty \quad \forall\theta \in \Theta$

iii) $H_\theta = \{y \in R^p:\ g(y;\theta) > 0\} = H$ independent of θ

iv) $(\partial/\partial\theta) \log g(y;\theta)$ is continuous in θ, $\forall y \in H$

v) For all n and all $y \in H$ there exists an unique $\theta \in \Theta$ noted $\hat{\theta}_n$ such that $K_n(\hat{\theta}_n) = 0$, where

$$K_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\partial/\partial\theta) \log f(X_i;\theta) + \frac{1}{n}(\partial/\partial\theta) \log g(y;\theta) \tag{2.8}$$

then $\hat{\theta}_n \to \theta_o$ almost surely when $n \to \infty$.

Proof: For any θ $(\partial/\partial\theta) \log g(y;\theta)$ is finite because of iv) so that $K_n(\theta)$ converges almost surely to the same expression as the likelihood equation based on the X_i's , i.e., by i) and Kolmogorov's theorem:

$$E_{\theta_o}\{(\partial/\partial\theta) \log f(X;\theta)\} = S(\theta) \tag{2.9}$$

Assumptions i) and ii) imply that $S(\theta_o) = 0$ and, for any ε small enough, $S(\theta_o - \varepsilon) > 0$ and $S(\theta_o + \varepsilon) < 0$. Thus for almost all sequences $K_n(\theta)$ there exists N such that for $n \geq N$

$$K_n(\theta_o - \varepsilon) > 0 \text{ and } K_n(\theta_o + \varepsilon) < 0 \quad .$$

Because of the continuity of $K_n(\theta)$ with respect to θ, its unique root $\hat{\theta}_n$ is in the interval $(\theta_o - \varepsilon, \theta_o + \varepsilon)$ and since ε can be chosen arbitrarily small $\hat{\theta}_n \to \theta_o$, with probability one.

Theorem 2.9: Under the assumptions of theorem (2.8) and the additional assumption that $(\partial/\partial\theta) \log f(x;\theta)$ is a monotone function of θ in some neighborhood of θ_o, for any $y_1, y_2 \in H$ we have when $n \to \infty$,

$$\frac{\hat{L}(y_1;\underset{\sim}{X})}{\hat{L}(y_2;\underset{\sim}{X})} = \frac{g(y_1;\hat{\theta}_n^1)}{g(y_2;\hat{\theta}_n^2)} * \frac{\prod_{i=1}^{n} f(X_i;\hat{\theta}_n^1)}{\prod_{i=1}^{n} f(X_i;\hat{\theta}_n^2)} \xrightarrow{a.s.} \frac{g(y_1;\theta_o)}{g(y_2;\theta_o)}$$

where $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ are the roots of $K_n(\theta)$ with $y = y_1$ and $y = y_2$ respectively.

Proof: If we can prove that the right ratio of the right hand side of the above equality tends a.s. to one or equivalently Z_n tends a.s. to zero, the theorem is proven since $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ tend to θ_o a.s. and g is a continuous function of θ, where

$$Z_n = \sum_{i=1}^{n}\{\log f(X_i;\hat{\theta}_n^1) - \log f(X_i;\hat{\theta}_n^2)\} .$$

We show that whenever the sequences $\{\hat{\theta}_n^1\}$ and $\{\hat{\theta}_n^2\}$ converge to θ_o it implies that Z_n tends to zero, and thus

$$\Pr\{Z_n \to 0\} \geq \Pr\{\hat{\theta}_n^1, \hat{\theta}_n^2 \to 0\} = 1 ,$$

i.e., Z_n tends to zero almost surely.

Using a Taylor expansion we can write

$$Z_n = (\hat{\theta}_n^1 - \hat{\theta}_n^2) \sum_{i=1}^{n}(\partial/\partial\theta) \log f(X_i;\theta_n^*) \tag{2.10}$$

where θ_n^* is between $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$. By definition $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ satisfy

$$\sum_{i=1}^{n}(\partial/\partial\theta) \log f(X_i;\hat{\theta}_n^k) = -(\partial/\partial\theta) \log g(y_k;\hat{\theta}_n^k) \quad k=1,2 \tag{2.11}$$

We can choose n large enough so that $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ be in a neighborhood of θ_o where $(\partial/\partial\theta) \log f(x;\theta)$ is monotone in θ. From iv) $(\partial/\partial\theta) \log g(y_k;\theta)$, $k = 1,2$, are bounded in this neighborhood and therefore the left hand sides of the equations (2.11) are also bounded. Since $\hat{\theta}_n^*$ lies between $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ it follows from the monotonicity that $\sum_{i=1}^{n}(\partial/\partial\theta)\log f(X_i;\theta_n^*)$ is bounded and thus from (2.10) $\hat{\theta}_n^1 - \hat{\theta}_n^2$ tends to zero implies Z_n tends to zero, which completes the proof.

Remarks 2.10

a) Because of the assumption of unicity in v) the maximum likelihood estimator based on $X_1, X_2, \ldots, X_n$ and y, when it exists, has to be $\hat{\theta}_n$. Its existence, however, is required in the definition of the PLF.

b) By picking up a fixed value for y_2 and setting $y = y_1$ a variation of theorem 2.9 is that there exists a function k such that $k(\underset{\sim}{X})\hat{L}(y;\underset{\sim}{X}) \to g(y;\theta_o)$ almost surely.

c) All previous results are true for the regular class of exponential families as defined by Zacks (1971), theorem 5.1.2.

d) Note also that if $(\partial/\partial\theta)\log g(y;\theta)$ is a bounded function of y for any fixed $\theta \in \Theta$, then from equality (2.8) the convergence of $K_n(\theta)$, and consequently that of $\hat{\theta}_n$, is uniform in y. Furthermore the summation in Taylor's expansion (2.10) is bounded and thus the convergence of Z_n is uniform in y as well. If additionally $g(y;\theta)$ is continuous in θ uniformly in y, then the convergence of the PLF is uniform in y.

e) For the regular class of exponential families, $h(y)\exp\{\theta y + k(\theta)\}$, the two conditions above are satisfied for y on any bounded interval which generally will be sufficient to have a uniform convergence for all y because density functions have to vanish at infinity (see proof in section III.6.1).

f) It seems that the previous results could be generalized to the case where θ is k-dimensional by taking assumptions similar to those for the classical proof of the consistency of the MLE (see Zacks (1971)).

Corollary 2.11: Under the assumptions of theorem (2.9), Fisher's likelihood tends to $g(y;\theta_o)/g(y;\hat{\theta}(y))$ a.s. when $n \to \infty$, where $\hat{\theta}(y)$ is the MLE based on y .

Proof: Consider the ratio $\tilde{L}(y;\underset{\sim}{X})/\tilde{L}(y_2;\underset{\sim}{X})$ for $y_1, y_2 \in H$. Then

$$\frac{\tilde{L}(y_1;\underset{\sim}{X})}{\tilde{L}(y_2;\underset{\sim}{X})} = \frac{R(\hat{\theta}_n^1;\underset{\sim}{X}) * R(\hat{\theta}_n^1;y_1)}{R(\hat{\theta}_n^2;\underset{\sim}{X}) * R(\hat{\theta}_n^2;y_2)} ,$$

where $\hat{\theta}_n^1$ and $\hat{\theta}_n^2$ are the same as previously. Therefore

$$\frac{\tilde{L}(y_1;\underset{\sim}{X})}{\tilde{L}(y_2;\underset{\sim}{X})} = \frac{\prod_{i=1}^{n} f(X_i;\hat{\theta}_n^1)}{\prod_{i=1}^{n} f(X_i;\hat{\theta}_n^2)} * \frac{g(y_1;\hat{\theta}_n^1)/g(y_1;\hat{\theta}(y_1))}{g(y_2;\hat{\theta}_n^2)/g(y_2;\hat{\theta}(y_2))} ,$$

which shows the relationship between Fisher's likelihood and the PLF. That is

$$\frac{\tilde{L}(y_1;\underset{\sim}{X})}{\tilde{L}(y_2;\underset{\sim}{X})} = \frac{L(y_1;\underset{\sim}{X})}{L(y_2;\underset{\sim}{X})} * \frac{g(y_2;\hat{\theta}(y_2))}{g(y_1;\hat{\theta}(y_1))} \tag{2.12}$$

But by theorem (2.8) the first ratio on the right hand side tends to $g(y_1,\theta_o)/g(y_2,\theta_o)$ a.s., so that

$$\frac{\tilde{L}(y_1;\underset{\sim}{X})}{\tilde{L}(y_2;\underset{\sim}{X})} \overset{\text{a.s.}}{\rightarrow} \frac{g(y_1;\theta_o)/g(y_1;\hat{\theta}(y_1))}{g(y_2;\theta_o)/g(y_2;\hat{\theta}(y_2))}$$

and the result stated follows from the normalization of $\tilde{L}(y;\underset{\sim}{X})$.

<u>Remark 1:</u> Suppose y is the sample mean (or total) of a sample of size m from the same population as $x_1,x_2,\ldots,x_n$. If for all θ the density functions satisfy the assumptions of the central limit theorem, then the density of y tends to be normal when $m \to \infty$. Thus when both m and n tend to infinity we expect the PLF to tend to have a normal shape.

<u>Remark 2:</u> One may ask about the behavior of $\hat{L}(y;x)$ when $m \to \infty$ and n remains finite. As will be shown in special cases ulteriorally, the limiting form depends on the distribution of the population.

<u>Remark 3:</u> Another question of interest is the existence of a Bayes prior yielding a Bayesian predictive density proportional to the PLF. We restrict our investigation to y being the sum of a future sample from the exponential family

$$f(x;\theta) = h(x)\, \exp\{\theta x + \psi(\theta)\}$$

i.e. the subclass of the exponential family for which the sample sum is a sufficient statistic.

Then denoting by x and y the sample sums, we have

$$\hat{L}(y;x) = k(x,n,m)\ \exp\{(x+y)\hat{\theta} + (n+m)\psi(\hat{\theta})\} \tag{2.13}$$

where $\hat{\theta}$ satisfies

$$x + y + (n+m)(\partial/\partial\theta)\psi(\theta) = 0$$

or

$$\hat{\theta} = \frac{\partial\psi}{\partial\theta}^{-1}[-(x+y)/(m+n)] \quad .$$

For a prior $\pi(\theta)$ we obtain the density of y ,

$$f_{\pi}(y|x) = h(x,m,n) \int_{\Theta} \exp\{(x+y)\theta + (n+m)\psi(\theta)\}\,\pi(\theta)d\theta$$

Thus we are looking for a solution π for the integral equation

$$\int_{\Theta} \exp\{t\theta + N\psi(\theta)\}\pi(\theta)d\theta = q(x,m,n)\ \exp\{t\hat{\theta} + N\psi(\hat{\theta})\} \quad \forall t,$$

where $t = x + y$ and $N = n + m$. But since $\hat{\theta}$ is a function of t only, the previous equation becomes

$$\int_{\Theta} \exp\{t\theta + N\psi(\theta)\}\,\pi(\theta)d\theta = c(N)\ \exp\{t\hat{\theta} + N\psi(\hat{\theta})\} \quad \forall t \ .$$

If Θ is the whole real line (resp. half real line) we see by setting $t = -s$ that a necessary and sufficient condition for π to exist is that the right hand side, as a function of t, be a bilateral (resp. one-sided) Laplace-Transform of a positive function, i.e.,

$$\pi(\theta) = c(N)\exp\{-t\theta - N\psi(\theta)\}\,\mathcal{L}^{-1}[\exp\{t\hat{\theta} + N\psi(\hat{\theta})\}] \;.$$

Generally the right hand side will be a rather complex function and no explicit solution will be available. The existence of an inverse will be easy to verify, but the conditions for it to be positive as given by Widder (1971) or Doetsch (1950) are hard to verify.

For the Poisson and Binomial cases we will exhibit, however, a Bayesian predictive density that is very close to the PLF.

II.4 Likelihood function and probability distribution.

In the classical likelihood terminology the likelihood ratio for two values of an unknown parameter is the ratio of the probabilities of observing the data x under both states of nature. The type of likelihoods we defined previously is conceptually different. A likelihood ratio for two values y_1 and y_2 is related to the ratio of the highest probabilities of observing y_1 and y_2 themselves, although in combination with the observed data x.

As for the parameter in the classical approach, the probability distribution of y cannot be recovered from its likelihood since we compare probabilities of events under different states of nature with $\hat{\theta}_{xy}$ a function of y.

Whereas in the classical case a bridge from likelihood to probability distribution of an unknown parameter is provided by the Bayesian argument, in prediction theory a bridge can be provided by the asymptotic property in theorem 2.9. Just as not any prior is valid for the Bayesian argument, here the PLF is not necessarily integrable; but the fact that it is integrable asymptotically indicates that it might very well be so for n finite, and in such a case the density function obtained by proportionality to the PLF defines a distribution that we name the "Maximum Likelihood Predictive Distribution" or briefly MLPD.

III SIMPLE RANDOM SAMPLING FROM A POISSON DISTRIBUTION

The methods presented in the preceding chapter are now applied to the Poisson distribution. The MLPD exists and it is shown how its CDF[1a] is located with respect to the CDF's derived from the other methods. The asymptotic behaviors are emphasized for n and/or m tending to infinity. An explicit expression of the density of the MLPD is not available but a quite accurate approximation is developed.

III.1 Problem and application.

Suppose we observe a sample of size n , $x_1, x_2, x_3, \ldots, x_n$, from a Poisson distribution with unknown parameter λ. What prediction interval can we give at a chosen confidence level for the sum of a future sample drawn from the same population?

We assume that we are sampling from an infinite population. However the case of a finite population can be brought into the same framework if we assume that this finite population has originated from a "superpopulation" with a Poisson distribution. Then we make the prediction on the total of the whole finite population in the following way:

We sample n items out of the N items constituting the whole population and thereby observe a total x for the sample. We now consider the remaining $N - n$ items with sum y as our future sample. After computing a prediction interval on y ,

$$Pr_\lambda\{L(x) \leq y \leq U(x)\} = \alpha \qquad \forall\lambda,$$

[1a] CDF = Cumulative Distribution Function

we derive a prediction interval for the total $t = x + y$,

$$\Pr_\lambda\{L(x) + x \leq t \leq U(x) + x\} = \alpha \qquad \forall\lambda \quad .$$

III.2 Frequentist conditional approach.

A sufficient (and complete) statistic for the joint distribution of X and Y is $T = X + Y$. It can easily be shown that the distribution of Y given T is Binomial with parameters T and $q = m/(m + n)$,

$$f_{Y|T}(y|t) = \binom{t}{y} q^y p^{t-y}, \quad y = 0,1,2,\ldots,t \quad .$$

Then from theorem 2.2 the conservative α-upper limit U is defined by the following inequalities:

$$\sum_{k=0}^{U} B(k;\ U + x + 1,q) \geq \alpha$$

and
$$\sum_{k=0}^{U-1} B(k;\ U + x,q) < \alpha \quad .$$

But it can be shown (see Olsen (1974).) that

$$\sum_{k=0}^{s} B(k;\ s + r,\theta) = \Pr\{Z \leq s\} \quad ,$$

where Z is a random variable having a $NB(r,\theta)$[2] distribution.

[2] Z is said to have a Negative Binomial distribution with parameters (r,θ) when $\Pr\{Z = z\} = \binom{z + r - 1}{z} \theta^r (1 - \theta)^Z \quad Z \in I = \{0,1,2,\ldots\}$.

Then $E(Z) = (1-\theta)r/\theta$ and $\psi_Z(t) = \left(\frac{\theta}{1 - (1-\theta)e^t}\right)^r$ (see Feller (1957).).

Consequently U is the α-upper fractile of a $NB(x + 1,p)$ distribution (see also corollary 2.3.).

For a β-lower limit we have from theorem 2.4

$$\sum_{k=0}^{L} B(k; L + x,q) > 1 - \beta$$

$$\sum_{k=0}^{L-1} B(k; L + x - 1,q) \leq 1 - \beta \quad .$$

That is L is the β-lower fractile of a $NB(x,p)$ distribution (see corollary 2.5), except for $x = 0$ where $L = 0$.

We summarize the previous results in the following theorem.

Theorem 3.1: The α-upper prediction limit given by the frequentist conditional approach for the sum of a random sample of size m from a Poisson distribution, given the sum x of a sample of size n , is the α-upper fractile of the $NB[x + 1, n/(n + m)]$ distribution. The β-lower prediction limit is the β-lower fractile of the $NB[x,n/(n + m)]$ distribution (except for $x = 0$).

III.3 Bayesian approach

In the second chapter we have seen that the Bayesian predictive density of Y is defined by

$$f(y\ x) = \frac{\int_0^\infty f_Y(y;\lambda) f_X(x;\lambda)\pi(\lambda)d\lambda}{\int_0^\infty f_X(x;\lambda)\pi(\lambda)d\lambda}$$

Consider priors of the form $\lambda^{-\alpha}$. Then we have

$$f_\alpha(y|x) = \frac{(x!y!)^{-1} \int_0^\infty e^{-m\lambda}(m\lambda)^y e^{-n\lambda}(n\lambda)^x \lambda^{-\alpha} d\lambda}{(x!)^{-1} \int_0^\infty e^{-n\lambda}(n\lambda)^x \lambda^{-\alpha} d\lambda}$$

where in order for both integrals to be convergent, for a given x, we must have $\alpha < x + 1$.

Then introducing the Gamma function we can write

$$f\ (y|x) = \frac{m^y}{y!} \frac{\Gamma(x + y + 1 - \alpha)/(m + n)^{x + y + 1 - \alpha}}{\Gamma(x + 1 - \alpha)/n^{x + 1 - \alpha}}$$

and further by setting $p = n/(n + m)$ and $q = m/(n + m)$,

$$f\ (y|x) = \frac{\Gamma(x + y + 1 - \alpha)}{\Gamma(x + 1 - \alpha)y!} p^{x + 1 - \alpha} q^y$$

for $y \varepsilon I = \{0,1,2,...\}$

In particular for a uniform prior, i.e. $\alpha = 0$, we have

$$Y|x \sim NB(x + 1,p) \quad .$$

For a $1/\lambda$ prior (assuming $x \neq 0$) we have

$$Y|x \sim NB(x,p) \quad .$$

The next theorem follows immediately.

Theorem 3.2: The frequentist lower and upper prediction limits as described in the theorem of section III.2 coincide respectively with the Bayes lower prediction limit under prior $1/\lambda$ (when it exists, i.e., $x \neq 0$) and the Bayes upper prediction limit under uniform prior.

Using an argument similar to the one given by Thatcher (1964) for the binomial prediction one could show that no prior will yield the same limits as the frequentist approach.

III.4 Likelihood approaches

III.4.1 Fisher's likelihood

Applying Fisher's likelihood approach to the Poisson problem gives

$$\tilde{L}(y;x) = R(x;\hat{\lambda})\ R(y;\hat{\lambda})$$

$$= \frac{e^{-n\hat{\lambda}}\hat{\lambda}^{x}}{e^{-x}(x/n)^{x}} * \frac{e^{-m\hat{\lambda}}\ (\hat{\lambda})^{y}}{e^{-y}(y/m)^{y}} \quad ,$$

with $\hat{\lambda} = (x + y)/(m + n)$, or

$$\tilde{L}(y;x) = (x + y)^{x + y} p^{x} q^{y}/(x^{x} y^{y}) \quad ,$$

with $p = n/(m + n)$, $q = m/(m + n)$.

The maximum 1 is obtained for $\bar{\tilde{y}} = \bar{x}$, i.e. $\tilde{y} = mx/n$.

III.4.2 Prediction likelihood approach (PLF)

The predictive likelihood approach yields

$$\hat{L}(y;x) \propto f_X(x;\hat{\lambda}) f_Y(y;\hat{\lambda}) \quad \text{with} \quad \hat{\lambda} = (x + y)/(m + n)$$

$$\propto e^{-(n + m)\hat{\lambda}} n^{x} m^{y} (\hat{\lambda})^{x + y}/(x!y!)$$

or

$$\hat{L}(y;x) \quad \propto \quad e^{-y}(x + y)^{x + y}\, q^y/y!$$

We now look for the value $\hat{y}$ of y that maximizes $\hat{L}(y;x)$. In order to do this we have to make certain approximations. First using Stirling's formula,

$$y! \; = \; \sqrt{2\pi}\;\alpha\; y^{y + \frac{1}{2}} e^{-y} \quad ,$$

we obtain

$$\hat{L}(y;x) \propto (x + y)^{x + y} q^y / y^{y + \frac{1}{2}} \quad .$$

Taking the logarithm of the right side we have

$$(x + y) \log (x + y) + y \log q - (y + \tfrac{1}{2}) \log y \quad ,$$

and by differentiation

$$\log (x + \hat{y}) + \log q - \log \hat{y} - \tfrac{1}{2}\hat{y} = 0$$

or $$(x + \hat{y})q/\hat{y} = e^{\frac{1}{2}\hat{y}} \; = \; 1 + \tfrac{1}{2}\hat{y} \; + \; \ldots$$

i.e. for $\hat{y}$ not small: $(x + \hat{y})q \simeq \hat{y} + \frac{1}{2}$

or $$\hat{y} \cong \frac{m}{n}(x - \tfrac{1}{2}) - \tfrac{1}{2} \quad .$$

We see that $\hat{y}$ is less than $\tilde{y} = mx/n$ of Fisher's approach.

III.4.3 Comparison of the two likelihoods

Let us look at the ratio of Fisher's likelihood to the PLF as a function of y, that is

$$\tilde{L}(y;x)/\hat{L}(y;x) \quad \propto \quad \frac{y^y}{e^{-y}\, q^y}$$

By Stirling's formula which is fairly accurate even for small values of y, the right hand side is approximately equal to $\sqrt{y}$. Thus the above ratio is an increasing function of y; it could be shown by using a theorem analogous to theorem 3.4 for likelihood functions, that Fisher's likelihood is always located to the right of the PLF. This is confirmed by the fact that $\tilde{y}$ is larger than $\hat{y}$.

More will be said about the two likelihoods when we study their asymptotic behavior. However we can already note that the maxima will tend to be the same only when $x \to \infty$, m/n not going to 0.

III.5 Comparison between the MLPD and the Bayesian predictive distributions.

We first need to establish theorem 3.3 and theorem 3.4. (see also Pratt (1965).).

Theorem 3.3: Let $g(x)$ be a density function strictly positive on the interval $[a,+\infty)$ where a is finite. If $f(x)$ is a function positive on the same interval and such that $f(x)/g(x)$ is non-increasing for $x > M$, then $f(x)$ can be normalized to a density function, i.e. $\int_a^{+\infty} f(t)dt < \infty$.

Proof:
$$\int_a^{+\infty} f(t)dt = \int_a^M f(t)dt + \int_M^{+\infty} f(t)dt$$

$$= K + \int_M^{+\infty} [f(t)/g(t)] \; g(t)dt$$

$$\leq K + [f(M)/g(M)] \int_M^{+\infty} g(t)dt$$

$$\leq K + [f(M)/g(M)] \, [1 - G(M)] < \infty$$

This theorem holds as well in the discrete case, if $f(x)$ and $g(x)$ are strictly positive on the same discrete set of values of x. The integral is then replaced by a summation sign. For convenience the interval of definition has been chosen as $[a,+\infty)$. For an interval $(-\infty,b]$ the convergence is insured when $f(x)/g(x)$ is non-decreasing.

<u>Theorem 3.4:</u> If $f(x)$ and $g(x)$ are density functions strictly positive on the same interval (a,b) and $f(x)/g(x)$ is a non-decreasing function on (a,b) then for their corresponding CDF's we have $F(x) \leq G(x)$.

<u>Proof</u>: For $x \leq a$, $F(x) = G(x) = 0$

For $x \geq b$, $F(x) = G(x) = 1$

For $a < x < b$,

$$F(x) = \int_a^x [f(t)/g(t)]g(t)dt \leq [f(x)/g(x)] \int_a^x g(t)dt$$

$$= [f(x)/g(x)]G(x)$$

$$1 - F(x) = \int_x^b [f(t)/g(t)]g(t)dt \geq f(x)/g(x) \int_x^b g(t)dt$$

$$= [f(x)/g(x)][1 - G(x)]$$

then:

$$\frac{F(x)}{1 - F(x)} \leq \frac{G(x)}{1 - G(x)} \Longrightarrow F(x) \leq G(x) \quad .$$

The proof holds for the discrete case when the integral sign is replaced by a summation sign.

We now come to the main statement of this section:

Theorem 3.5: When random sampling from a Poisson distribution there exists a Maximum Likelihood Predictive Distribution, noted $\hat{F}(y;x)$, for the sum, Y, of a future sample. Moreover if we denote by $F_1(y|x)$ and $F_2(y|x)$ the Bayesian predictive CDF's for a $1/\lambda$ and a uniform prior respectively the following inequalities hold:

$$F_2(y|x) \leq \hat{F}(y;x) \leq F_1(y|x)$$

Proof: Let $\hat{L}(y;x)$ be the PLF and $f_1(y|x)$ and $f_2(y|x)$ the probability mass functions for the Bayesian $1/\lambda$ and uniform priors. We are going to show that:

i) $\hat{L}(y;x)/f_2(y|x)$ is a non-increasing function of y for all $x \varepsilon \{I = 0,1,2,\ldots\}$. Then by theorem 3.3 the existence of $\hat{F}(y;x)$ is proven and by theorem 3.4 the left inequality holds.

ii) $\hat{L}(y;x)/f_1(y|x)$ is non-decreasing in y and by theorem 3.4 the right inequality holds.

Recall that

$$\hat{L}(y;x) = k(x_1,n,m)\ e^{-(x + y)}(x + y)^{x + y}q^y/y! \quad ,$$

$$f_2(y\ x) = k'(x,n,m)\ (x + y)!\ q^y/y! \quad ,$$

and $$f_1(y|x) = k''(x,n,m)\ (x + y - 1)!\ q^y/y! \quad .$$

For i) we look at the ratio

$$\hat{L}(y;x)/f_2(y|x) = h(x,m,n)\ e^{-(x + y)}(x + y)^{x + y}/(x + y)!$$

and have to show equivalently that $u_n = e^{-n}n^n/n!$
is a non-increasing sequence. Let $v_n = \log u_n$. Then

$$v_{n+1} - v_n = -(n+1)\log(n+1) - \log(n+1)! + n - n\log n + \log n!$$

$$= -1 + n\log(1+1/n) \qquad (n \neq 0)$$

But since $\log(1+x) < x$ for all $x \neq 0$ we have

$$v_{n+1} - v_n < -1 + n(1/n) \quad \text{or} \quad v_{n+1} - v_n < 0 \quad \text{for} \quad n = 1,2,...$$

Consequently $u_{n+1} < u_n$ for $n = 1,2,...$ and this holds also for $n = 0$ since $u_0 = 1$ and $u_1 = 1/e$.

For ii) we consider:

$$\hat{L}(y;x)/f_1(y|x) = h^1(x,m,n)\ e^{-(x + y)}(x+y)^{x+y}/(x+y-1)!$$

i.e. we have to show that $w_n = e^{-n}n^n/(n - 1)!$
in a non-decreasing sequence. Note that the sequence is not defined for $n = 0$ which corresponds to the fact that the prior $1/\lambda$ does not apply when $x = 0$.

$$s_n = \log w_n = -n + n\log n - \log(n - 1)!$$

$$s_{n+1} - s_n = (n+1)\log(1+1/n) - 1$$

But since $\log x > 1 - 1/x$ for $x > 1$, taking $x = 1 + 1/n$ we find

$$s_{n+1} - s_n > 0 .$$

We will see in section II.6.2 that $\tilde{L}(y;x)$ can also be normalized to a density and that the corresponding CDF is always below $F_2(y|x)$.

<u>Existence of a Bayes prior yielding the MLPD:</u>

At this point we may ask if there is a Bayes prior $\pi(\lambda,n,m)$ that leads to $\hat{f}(y;x)$, i.e. that satisfies

$$(x!y!)^{-1} \int_0^\infty e^{-(n+m)\lambda} (n\lambda)^x (m\lambda)^y \, (\lambda,n,m) d\lambda = \tag{3.1}$$

$$k(x,n,m) e^{-(x+y)} (x+y)^{x+y} q^y / y!$$

for $x,y \in I = \{0,1,2,\ldots\}$.

By setting $t = x + y$ and $N = m + n$, (3.1) can be rewritten equivalently as

$$\int_0^\infty e^{-N\lambda} \lambda^t \pi(\lambda,m,n) d\lambda = k(x,m,n) \frac{e^{-t} t^t}{N^t} , \quad \text{for all } t, x \in I$$

This equality shows that k does not depend on x , and further, π depends only on N and λ. Then,

$$\int_0^\infty e^{-N\lambda} (N\lambda)^t \pi(\lambda,N) d\lambda = e^{-t} t^t \quad \text{for all } t \in I$$

By setting $u = N\lambda$ one sees that $\pi(\lambda;N)$ has to be of the form

$N^{-1}\psi(N\lambda)$ and an equivalent condition for the existence of a Bayes prior is the existence of a non-negative function p such that

$$\int_0^\infty e^{-u} u^t p(u)du = e^{-t} t^t \qquad t \in I \tag{3.2}$$

The conditions for $e^{-t}t^t$ to be a bilateral Laplace transform (see section II.3.2) are fulfilled and thus there exists a unique function $p(u)$ verifying (3.2). However $p(u)$ is not a classical function and we are unable to verify its positiveness.

Instead of this we exhibit a prior that leads to a predictive distribution that is very close to the MLPD. The prior we chose is

$$p(u) = u^{-\frac{1}{2}} \quad .$$

Using the formulae (see for example Widder (1961).)

$$\Gamma(t + \tfrac{1}{2}) = \sqrt{\pi}\ (2t)!2^{-2t}/t! \tag{3.2a}$$

and $$t! = \sqrt{2\pi}\ t^{t + \frac{1}{2}}e^{-t}\exp\{-1/12t) - 1/(360t^3) + o(t^{-5})\} \quad ,$$

we get

$$\Gamma(t + \tfrac{1}{2}) = \sqrt{2\pi}\, e^{-t}t^t \exp\{R(t)\} \tag{3.3}$$

with $$R(t) = -1(24t) + 7/(8 \times 360t^3) + o(t^{-5}) \quad .$$

$\Gamma(t + \frac{1}{2})$ gives a good a good approximation for $e^{-t}t^t$ and consequently so does the Bayesian distribution with prior $\lambda^{-\frac{1}{2}}$ for the MLPD. The larger $t = x + y$ the better the approximation. The posterior distribution for this prior is a $NB(x + 1/2,p)$[3] distribution.

Determination of the normalization constant for $\hat{f}(y;x)$

Our purpose is now to find the function $K(x,p)$ such that

$$K(x,p) \sum_{y=0}^{\infty} e^{-y}(x + y)^{x + y} q^y / y! = 1 \quad .$$

Consider the expression Q , where

$$Q = \frac{p^{x + \frac{1}{2}}}{x^x} \sum_{y=0}^{\infty} \frac{e^{-y}(x + y)^{x + y} q^y}{y!}$$

$$= \sum_{y=0}^{\infty} \frac{e^{-(x + y)}(x + y)^{x + y}}{e^{-x}x^x} \frac{p^{x + \frac{1}{2}} q^y}{y!}$$

We substitute for $e^{-(x + y)}(x + y)^{x + y}$ expressions from (3.3) and obtain

$$Q = \sum_{y=0} \frac{\Gamma(x + y + \frac{1}{2})}{\Gamma(x + \frac{1}{2})} \frac{p^{x + \frac{1}{2}} q^y}{y!} \exp\{R(x) - R(x + y)\} ,$$

or equivalently $Q = E[\exp\{R(x) - R(x + y)\}]$ where the expectation is to be taken with respect to a $NB(x + 1/2,p)$ distribution. But since

---

[3]We can extend the definition of a $NB(r,\theta)$ of section III.2 to non-integer values of r as follows:

$$\Pr\{Z = z\} = \frac{\Gamma(z + r)}{\Gamma(r)z!} \theta^r (1 - \theta)^z \qquad z \in I$$

$$\exp\{R(x) - R(x+y)\} \cong 1 - \frac{1}{24x} + \frac{1}{24(x+y)}$$

$$Q \cong 1 - \frac{1}{24x} + \frac{1}{24} E[\frac{1}{x+y}]$$

As a second order approximation we may compute the expectation with respect to a $NB(x+1,p)$ distribution:

$$E[\frac{1}{x+y}] \cong \sum_{y=0}^{\infty} \frac{(x+y)}{x!y!} p^{x+1} q^{y} \frac{1}{x+y}$$

$$\cong \frac{p}{x} \sum_{y=0}^{\infty} \frac{(x+y-1)!}{(x-1)!\, y!} p^{x} q^{x} = \frac{p}{x}$$

Thus $$Q \cong 1 - \frac{1}{24x} + \frac{1}{24}\frac{p}{x}$$

$$Q \cong 1 - \frac{q}{24x}$$

Going back to the definition of Q we establish formula (3.4):

$$(1 + \frac{q}{24x}) \frac{p^{x+\frac{1}{2}}}{x^{x}} \sum_{y=0}^{\infty} e^{-y}(x+y)^{x+y} q^{y}/y! \doteq 1 \qquad (3.4)$$

<u>Note 1:</u> Had we used for $\exp\{R(x)\}$ its exact value (see (3.3))

$$\exp\{R(x)\} = \Gamma(x+\tfrac{1}{2})/(\, 2\pi\, e^{-x}x^{x}) = (2x)!/(2^{2x+\frac{1}{2}}x!\,e^{-x}x^{x}),$$

we would have obtained (3.5) alternately to (3.4):

$$(1 - \frac{p}{24x}) \frac{2^{2x+\frac{1}{2}} x!\, p^{x+\frac{1}{2}}}{(2x)!} \sum_{y=0}^{\infty} e^{-(x+y)}(x+y)^{x+y} q^{y}/y! \doteq 1 \qquad (3.5)$$

The difference between the two approximations is negligible. Tables I and II give the exact value of the left hand side of (3.4) and (3.5) respectively for various values of x and p.

Note: Formulae (3.4) and (3.5) do not apply when $x = 0$. We will develop a solution which has been inspired by an article of Haight and Breuer (1960) on the Borel-Tanner distribution. This solution for $x = 0$ leads eventually to the general solution for the sum

$$S_x = \sum_{y=0}^{\infty} (y + x)^{y + x} e^{-y} q^y / y!$$

Borel (1942) has shown that defining $\beta = \alpha e^{-\alpha}$ we have the relationship

$$\alpha = \sum_{y=1}^{\infty} y^{y-1} \beta^y / y! \quad ,$$

which by taking $\beta = q/e$ and $\alpha = 1 - u$ can be written as

$$1 - u = \sum_{y=1}^{\infty} y^{y-1} e^{-y} q^y / y! \qquad \text{with} \quad \underline{q = (1 - u)e^u} \quad .$$

Let us differentiate this equality with respect to q, then

$$-du/dq = (1/q) \sum_{y=1}^{\infty} y^y e^{-y} q^y / y!$$

and by replacing $dq/du = - ue^u = -uq/(1 - u)$,

$$\sum_{y=1}^{\infty} y^y e^{-y} q^y / y! = (1 - u)/u \quad .$$

Finally we obtain

$$S_o = \sum_{y=0}^{\infty} y^y e^{-y} q^y / y! = 1/u \quad .$$

Furthermore by differentiating S_o with respect to q we have

$$\sum_{y=1}^{\infty} y^y e^{-y} q^{y-1} / (y-1)! = dS_o/dq$$

i.e.,

$$\begin{aligned} S_1 &= e(dS_o/dq) \\ &= e(du/dq)(dS_o/du) \\ &= (e/q)\ (1 - 1/u)\ (dS_o/du) \end{aligned}$$

and in the same way

$$S_{k+1} = (e/q)(1 - 1/u)(dS_k/du) \ , \text{ with } S_o = 1/u$$

or by setting $t = 1/u$

$$S_{k+1} = (e/q)(t^3 - t^2)\ (dS_k/dt) \tag{3.6}$$

with

$$S_o = t \quad \text{and} \quad \underline{q = (1 - 1/t)e^{1/t}} \ .$$

The relationship (3.6) allows us to obtain the expression of S_k by recurrence, e.g.

$$S_1 = (e/q)(t^3 - t^2) \quad .$$

TABLE I. Approximation (3.4) - (Value of left hand side)

P \ X =	1	2	3	4	5	6	7	8	9	10
.99	.99731	.99954	.99984	.99992	.99995	.99997	.99998	.99998	.99999	.99999
.95	.99790	.99971	.99992	.99997	.99999	.99999	1.00000	1.00000	1.00000	1.00000
.90	.99659	.99992	1.00001	1.00002	1.00002	1.00002	1.00001	1.00001	1.00001	1.00001
.80	.99988	1.00028	1.00018	1.00012	1.00008	1.00006	1.00005	1.00004	1.00003	1.00002
.70	1.00100	1.00058	1.00032	1.00020	1.00013	1.00009	1.00007	1.00005	1.00004	1.00004
.60	1.00193	1.00081	1.00042	1.00025	1.00017	1.00012	1.00009	1.00007	1.00005	1.00004
.50	1.00265	1.00097	1.00048	1.00028	1.00018	1.00013	1.00010	1.00007	1.00006	1.00005
.40	1.00312	1.00103	1.00049	1.00029	1.00019	1.00013	1.00010	1.00007	1.00006	1.00005
.30	1.00326	1.00100	1.00046	1.00026	1.00017	1.00012	1.00008	1.00006	1.00005	1.00004
.20	1.00298	1.00084	1.00037	1.00021	1.00013	1.00009	1.00006	1.00005	1.00004	1.00003
.10	1.00210	1.00052	1.00022	1.00012	1.00007	1.00004	1.00003	1.00002	1.00001	1.00001
.05	1.00130	1.00030	1.00012	1.00007	1.00004	1.00003	1.00002	1.00001	1.00001	1.00001

TABLE II. Approximation (3.5) - (Value of left hand side)

P \ X =	1	2	3	4	5	6	7	8	9	10
.99	1.00013	1.00004	1.00002	1.00001	1.00001	1.00001	1.00000	1.00000	1.00000	1.00000
.95	1.00064	1.00019	1.00009	1.00005	1.00004	1.00003	1.00002	1.00001	1.00001	1.00001
.90	1.00125	1.00036	1.00018	1.00010	1.00007	1.00005	1.00004	1.00003	1.00002	1.00002
.80	1.00236	1.00069	1.00033	1.00019	1.00012	1.00009	1.00006	1.00005	1.00004	1.00003
.70	1.00331	1.00095	1.00044	1.00025	1.00016	1.00012	1.00009	1.00007	1.00005	1.00004
.60	1.00407	1.00114	1.00052	1.00030	1.00019	1.00013	1.00010	1.00008	1.00006	1.00005
.50	1.00461	1.00125	1.00056	1.00032	1.00020	1.00014	1.00010	1.00008	1.00006	1.00005
.40	1.00490	1.00128	1.00056	1.00031	1.00020	1.00014	1.00010	1.00008	1.00006	1.00005
.30	1.00487	1.00120	1.00051	1.00028	1.00018	1.00012	1.00009	1.00007	1.00005	1.00004
.20	1.00443	1.00100	1.00041	1.00022	1.00013	1.00009	1.00006	1.00005	1.00004	1.00003
.10	1.00338	1.00064	1.00024	1.00012	1.00007	1.00005	1.00003	1.00002	1.00002	1.00001
.05	1.00249	1.00039	1.00013	1.00006	1.00003	1.00002	1.00001	1.00001	1.00000	1.00000

III.6 Asymptotic properties

III.6.1 Limits when $m/n \to 0$

This means that $n \to \infty$ and m remains finite, i.e., the observed sample size alone becomes large. The behavior of $\tilde{L}(y;x)$ and $\hat{L}(y;x)$ has been studied from a general point of view in chapter II. If λ_o is the true value of the parameter then $x/n \overset{a.s.}{\to} \lambda_o$ and further

$$\hat{f}(y;x) \overset{a.s.}{\to} f_Y(y;\lambda_o) = \frac{e^{-m\lambda_o}(m\lambda_o)^y}{y!}$$

$$\tilde{L}(y;x) \overset{a.s.}{\to} \frac{e^{-m\lambda_o}(m\lambda_o)^y}{e^{-y}\, y^y} \qquad \text{for} \quad y = 0,1,2,\ldots$$

which exhibits for the Poisson case the deceiving behavior of Fisher's likelihood.

As for the Bayesian predictions we know (see Feller (1957).) that if $Z \sim NB(r,\theta)$ and $r \to \infty$ but $r(1-\theta) \to \lambda_o$ then $Z \overset{d}{\to} P(\lambda_o)$. In the present case for a $NB(x,p)$ where $p = n/(m+n)$, with probability one $x \to \infty$ since $x/n \to \lambda_o$, but $x(1-p) = xm/(m+n) \to m\lambda_o$. The same is true for a $NB(x+1,p)$.

Thus the Bayesian predictive distributions under priors $1/\lambda$ and 1 also converge to the true distribution of Y almost surely.

Uniform convergence of $\hat{f}(y;x)$.

Let us show that the a.s. convergence of $\hat{f}(y;x)$ is uniform in y. Recall from theorem (2.9) that the convergence will be uniform in y

on a bounded interval if $f(y;\lambda_o)$ is continuous at λ_o uniformly in y on that interval. This is obviously true for an interval $I_K = \{0,1,2,\ldots,K\}$ where K is arbitrary, i.e., for any given η and ε one can find $N_1(K)$ such that for all $y \varepsilon K$

$$N \geq N_1 \Rightarrow \Pr\{ \sup_{n \geq N} |\hat{f}_n(y;x) - f(y;\lambda_o)| < \varepsilon/2\} \geq 1 - \eta .$$

Let us choose K such that $f(K;\lambda_o) < \varepsilon/2$. The fact that $\hat{f}_n(y;x)$ lies within $\varepsilon/2$ of $f(y;\lambda_o)$ implies, for ε chosen small enough, that for $y = K$ we are to the right of the maximum of $\hat{f}_n(y;x)$ and then $\hat{f}_n(y;x)$ is decreasing in Y, i.e., $\hat{f}_n(y;x) < \hat{f}_n(K;x) < \varepsilon$ for all $y > K$. Thus

$$|\hat{f}_n(y;x) - f(y;\lambda_o)| < \varepsilon/2 \quad \text{for } y \varepsilon I_k \Rightarrow |\hat{f}_n(y;x) - f(y;\lambda_o)| < \varepsilon$$
$$\text{for all } y .$$

Consequently

$$\Pr\{ \sup_{n \geq N} |\hat{f}_n(y;x) - f(y;\lambda_o)| < \varepsilon\} > 1 - \eta \quad \text{for } N \geq N_1 ,$$

where N_1 is independent of y and ε,η can be chosen arbitrarily small.

This proof applies more generally to any distribution as long as the density of y is continuous at the true value of the parameter uniformly in y on any bounded interval, and the density of the MLPD tends to zero in a monotone way when y tends to infinity.

III.6.2 Limits when m and $n \to \infty$, m/n remaining constant.

Let us recall first (see Johnson and Kotz (1969).) that if Z is distributed $NB(r,\theta)$, when $r \to \infty$ Z tends to follow a Normal distribution with mean and standard deviation

$$E(Z) = r(1-\theta)/\theta \quad , \quad \sigma(Z) = \sqrt{r(1 -)\theta}/\theta \ .$$

Thus for a Bayes $1/\lambda$ prior with probability 1 (since $x \to \infty$ with prob. 1) $Y|x$ tends to be Normal with mean

$$\begin{aligned} E(Y|x) &= x(1-p)/p \\ &= mx/n \end{aligned}$$

and standard deviation

$$\begin{aligned} \sigma(Y|x) &= \sqrt{x(1-p)}\ /p \\ &= \sqrt{mx(m+n)/n^2} \\ &= m\sqrt{\frac{x}{n}\left(\frac{1}{m}+\frac{1}{n}\right)} \ . \end{aligned}$$

This is equivalent to the result one would obtain by using the normal approximation for the distributions of the sample sums.

For the Bayes uniform prior, x is replaced by $x + 1$ which leads to the same limiting form. The MLPD is between these two Bayesian predictive distributions and thus has the same asymptotic behavior.

We show now that Fisher's likelihood can be standardized to a density whose limit is the same as above.

Recall that

$$\tilde{L}(y;x) \propto (x + y)^{x + y} q^y / y^y \qquad \text{for} \qquad y = 0,1,2,\ldots$$

Consider the function defined as

$$h(y;x) \propto \begin{cases} 0 & \text{for} \quad y = 0 \\ (x + y)^{x + y} q^y / (y - 1)! & \text{for} \quad y = 1,2,\ldots \end{cases}$$

Since $(y - 1)!/y^y$ is non-increasing, $\tilde{L}(y;x)/h(y;x)$ is non-increasing as well, so we are done with the proof if we can show that $\sum_{y=0}^{\infty} h(y;x)$ is convergent.

It is easily seen that the series

$$u_y = (x + y)^{x + y} q^y / (y - 1)! \qquad y = 1,2,\ldots$$

is convergent for all x because $u_y = v_k$, where

$$v_k = (x + 1 + k)^{x + 1 + k} q^{k + 1} / k! \qquad k = 0,1,2,\ldots$$

and v_k is proportional to $\hat{f}(k;\ x + 1)$.

Thus $h(y;x)$ can be defined as a density function and

$$h(y;x) = \hat{f}(\ y - 1;\ x + 1) \qquad \text{for} \quad y \geq 1 .$$

Consequently a density can be derived from Fisher's approach which we denote $\tilde{f}(y;x)$, and we have

$$\tilde{F}(y;x) = \sum_{k=0}^{y} \tilde{f}(k;x) \geq \sum_{k=0}^{y} h(y;x) = \sum_{k=0}^{y-1} \hat{f}(k;x + 1) = \hat{F}(y - 1;\ x + 1)$$

i.e.: $\tilde{F}(y;x) \geq \hat{F}(y;x+1) - \hat{f}(y;x+1)$.

But since for all y, $\hat{f}(y;x+1)$ tends to zero when $x \to \infty$ we have

$$\tilde{F}(y;x) \geq \hat{F}(y;x+1) - \varepsilon$$

where ε can be chosen as small as we desire.

Consider now the ratio

$$\tilde{f}(y;x)/f_2(y|x) \propto [(x+y)^{x+y}q^y/y^y]/[(x+y)!\ q^y/y!]$$

$$\propto \frac{(x+y)^{x+y}\ y!}{(x+y)!\ y^y} = w_y$$

The difference $\log w_{y+1} - \log w_y = (x+y)\log(1 + 1/(x+y)) - y\log(1+1/y)$

is always non-negative because the function $u\log(1+1/u)$ is increasing for u positive. Thus the above ratio is a non-increasing function of y and $\tilde{F}(y;x)$ is always below $F_2(y|x)$.

Finally we have

$$\hat{F}(y;x+1) - \varepsilon \leq \tilde{F}(y;x) \leq F_2(y|x)$$

When $x \to \infty$, $\hat{F}(y;x+1)$ and $F_2(y|x)$ have the same limiting form as $\hat{F}(y;x)$, and so does $\tilde{F}(y;x)$ since ε goes to zero. We summarize these results in the following theorem.

<u>Theorem 3.6</u>: For $n,m \to \infty$, m/n remaining constant, the predictive CDF's of Y: $F_1(y|x)$, $F_2(y|x)$, $\tilde{F}(y;x)$ and $\hat{F}(y;x)$ tend to be such that

$$\frac{Y/m - \bar{x}}{\sqrt{\bar{x}(\frac{1}{n}+\frac{1}{m})}} \xrightarrow{d} N(0,1)$$

with probability one.

III.6.3 Limits when $m \to \infty$, n remaining finite.

One can imagine situations where the future sample to be predicted would have a large size m . This for instance would be the case for the type of inference discussed in section III.1 when sampling from a finite but "large" population, i.e., in fact the sampling fraction $n/(n + m)$ has to be small. It might be of interest then to study the limiting form of the distribution of Y in order to approximate it. Also the prediction intervals obtained from a limiting predictive distribution of Y/m can be seen as confidence intervals on the true parameter λ_o .

i) Limit of $\hat{f}(y;x)$

$$\hat{f}(y;x) \propto \frac{e^{-y}(x + y)^{x + y}}{y!} \left(\frac{m}{m + n}\right)^{y}$$

When $m \to \infty$ then $E(Y;x) \to \infty$; thus we can use Stirling's approximation for $y!$ and obtain

$$\hat{f}(y;x) \propto y^{x - \frac{1}{2}}(1 + x/y)^{x + y}(1 + n/m)^{-y} \quad .$$

Let us look rather at the distribution of $\bar{Y} = Y/m$ because $E(\bar{Y};x)$ is finite. Then

$$\hat{f}(\bar{y};x) \propto \bar{y}^{x - \frac{1}{2}} \left(1 + \frac{x}{m\bar{y}}\right)^{x + m\bar{y}} (1 + n/m)^{-m\bar{y}}$$

When $m \to \infty$ we obtain

$$\hat{f}(\bar{y};x) \to k(x)\ \bar{y}^{x - \frac{1}{2}} e^{x} e^{-n\bar{y}} \propto \bar{y}^{x - \frac{1}{2}} e^{-n\bar{y}}$$

so that $\hat{f}(\bar{y};x) \to \mathrm{GAMMA}(x - \frac{1}{2}\ , 1/n)$.

Thus for m "large" we can approximate $\hat{f}(y;x)$ by a $\mathrm{GAMMA}(x - \frac{1}{2}\ , m/n)$ distribution

ii) Limit of $\tilde{f}(y;x)$

$$\tilde{f}(y;x) \propto \frac{(x + y)^{x + y} q^{y}}{x^{x} y^{y}}$$

$$\propto \ y^{x}(1 + x/y)^{x + y}(1 + n/m)^{-y} \quad .$$

In the same way as in i) we find that

$$\tilde{f}(\bar{y};x) \to \ \mathrm{GAMMA}(x,1/n)$$

iii) Limit of the Bayesian predictions

For $F_1(y|x)$ as well as $F_2(y|x)$, $m \to \infty$ corresponds to having the parameter θ of the Negative Binomial distribution tending to 0 .

Lemmma 3.7: Let Z be a random variable distributed $NB(r,\theta)$. When $\theta \to 0$ the random variable θZ tends in distribution to a $\mathrm{GAMMA}(r - 1,1)$ distribution.

Proof: The characteristic function of Z is (see section III.2)

$$\psi_Z(t) = e^{tr}\theta^r/[1 - (1 - \theta)e^t]^r ,$$

then for θZ, $\psi_{\theta Z}(t) = e^{\theta tr}\theta^r/[1 - (1-\theta)e^{\theta t}]^r$

When $\theta \to 0$ the denominator is equivalent to

$$[1 - (1 - \theta)(1 + \theta t + o(\theta^2)]^r = [\theta(1 - t) + o(\theta^2)]^r$$

or

$$\theta^r(1 - r)^r$$

Thus $\psi_{\theta Z}(t) \to (1 - t)^{-r}$ when $\theta \to 0$

which is characteristic function of a GAMMA$(r - 1,1)$.

Applying this result to the Bayesian case with $1/\lambda$ prior, we have

$$Y \sim NB(x,n/(n + m))$$

$$nY/(m + n) \xrightarrow{d} \text{GAMMA}(x - 1,1) \quad \text{when} \quad m \to \infty ,$$

or equivalently

$$nY/m \xrightarrow{d} \text{GAMMA}(x - 1,1) .$$

For a uniform prior we obtain

$$nY/m \xrightarrow{d} \text{GAMMA}(x,1)$$

These results are equivalent to the posterior densities of the unknown parameter λ for the same corresponding priors.

We summarize the results for m "large" in the table below. Recall that m "large" means in fact m/n large since all predictive densities considered depend solely on this ratio.

Approximation for m large

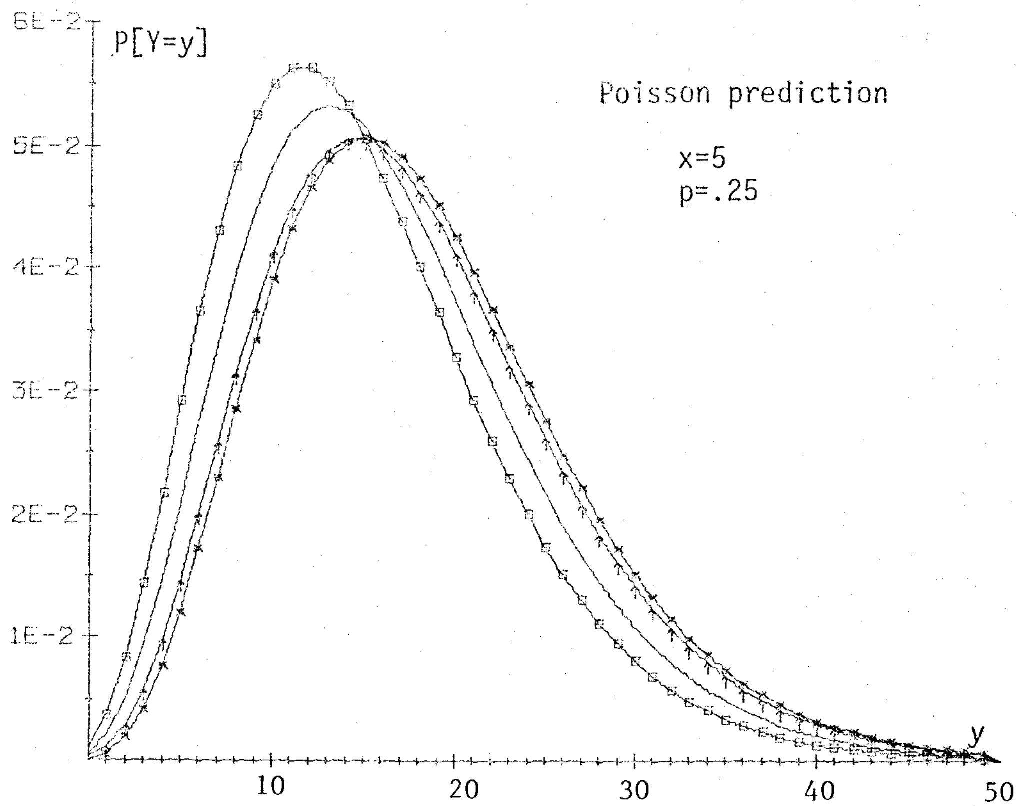
Type of Approach	Notation	Limiting Distribution	Limiting Expectation
Bayesian $1/\lambda$	$f_1(y\|x)$	$\Gamma(x - 1 ; m/n)$	$\frac{m}{n} x$
Bayesian 1	$f_2(y\|x)$	$\Gamma(x ; m/n)$	$\frac{m}{n}(x + 1)$
MLPD	$\hat{f}(y;x)$	$\Gamma(x - \frac{1}{2} ; m/n)$	$\frac{m}{n}(x + \frac{1}{2})$
FISHER	$\tilde{f}(y;x)$	$\Gamma(x ; m/n)$	$\frac{m}{n}(x + 1)$

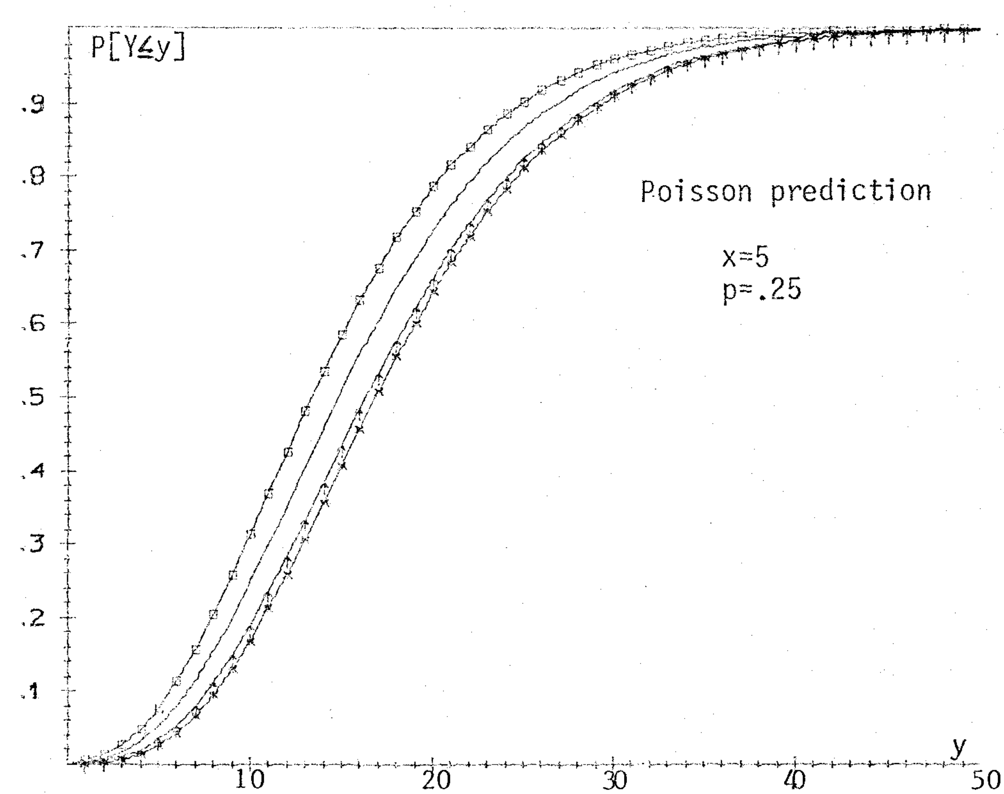
III.7 Examples:

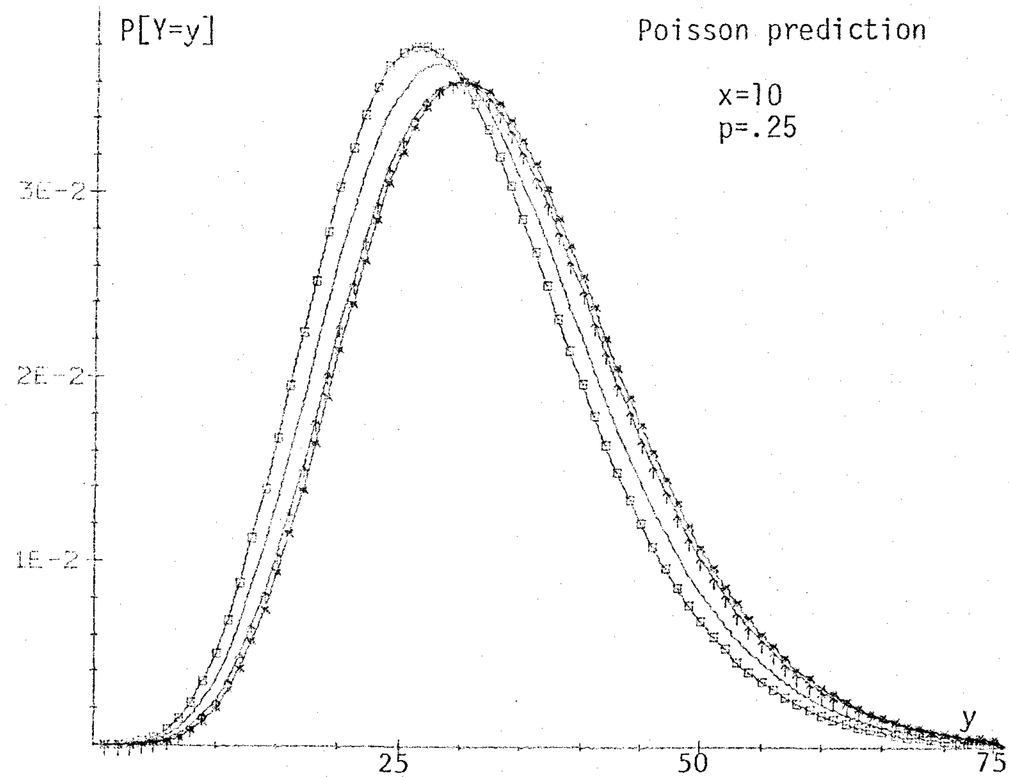
Several examples in the next pages illustrate the preceding results.

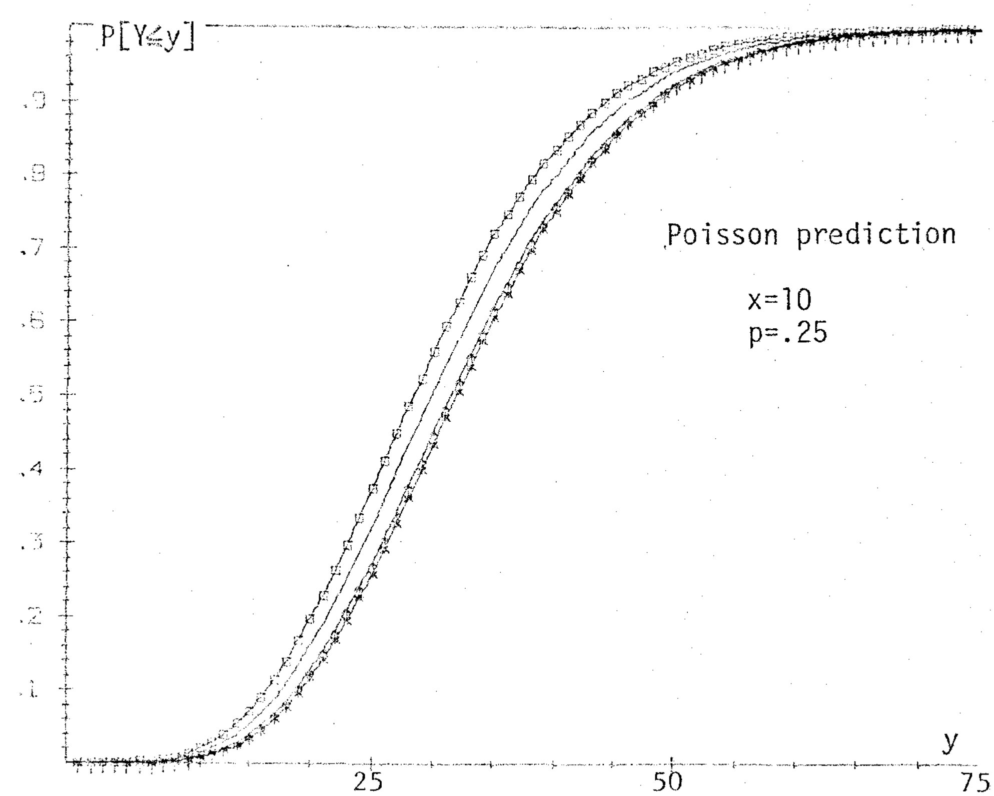
For a given value of x and $p = n/(n + m)$ the predictive density functions and cumulative distribution functions are plotted. These correspond respectively, from the left to the right, to the $NB(x,p)$, the MLPD, the $NB(x + 1,p)$ and the distribution derived from Fisher's approach.

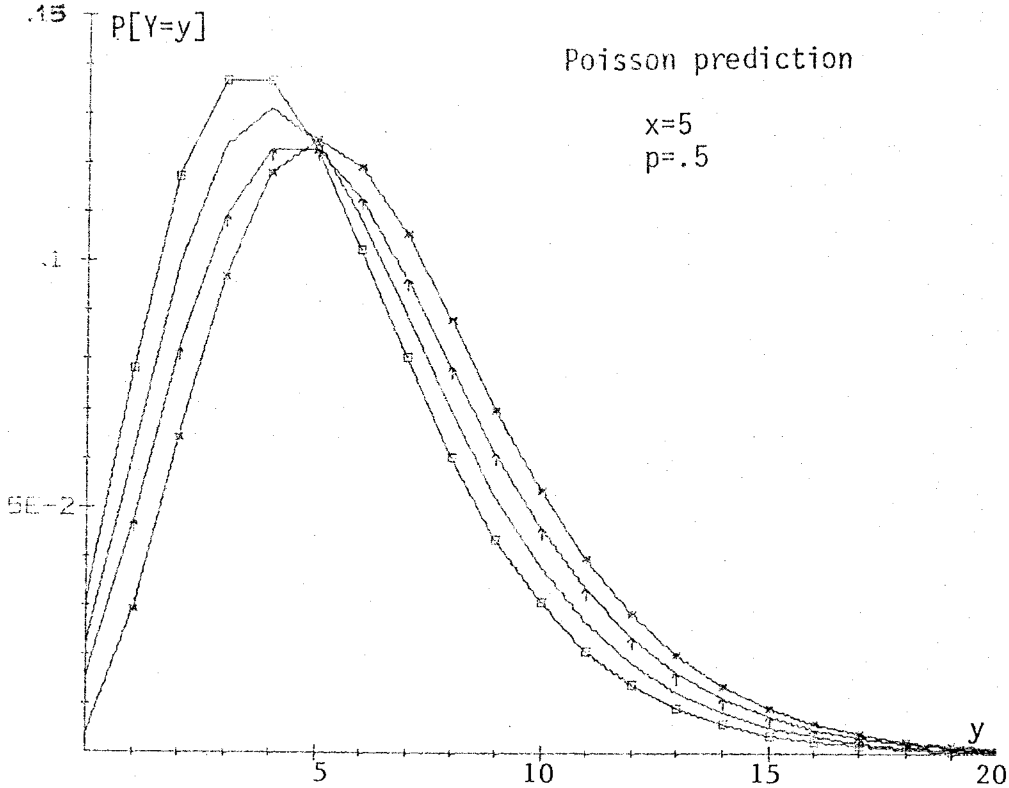
Six examples have been chosen, i.e., $x = 5$ and 10 for the values of p: .25, .50, .75 which correspond to n/m equal to $1/3$, 1 and 3.

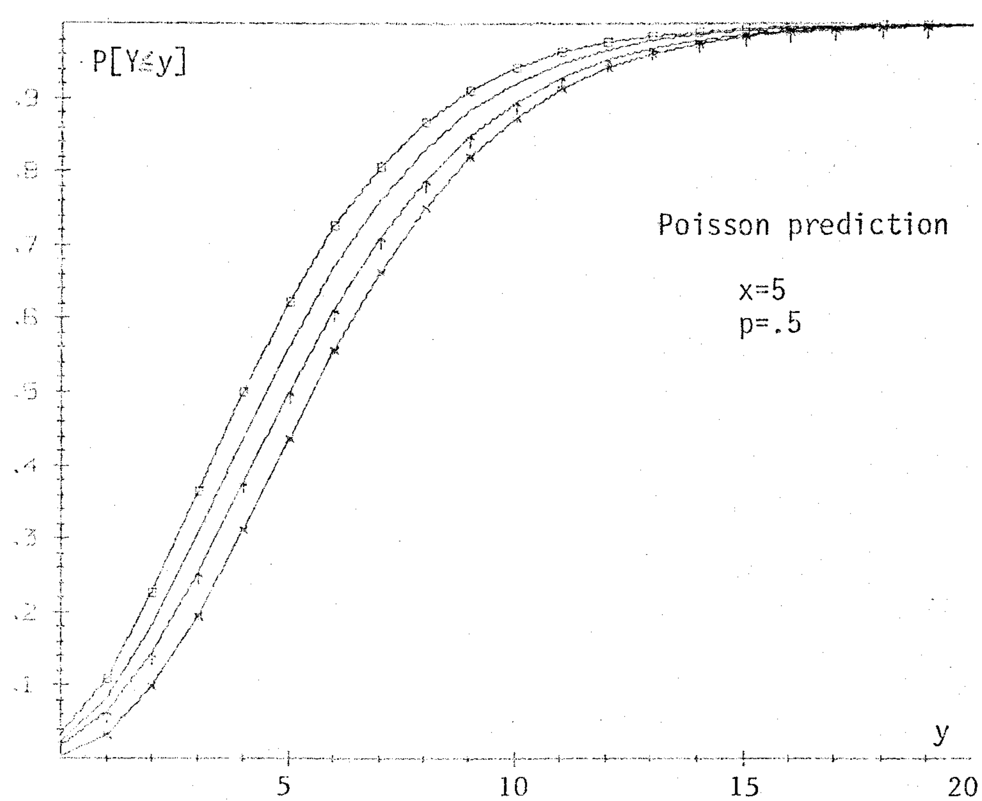

6E-2
5E-2
4E-2
3E-2
2E-2
1E-2
P[Y=y]
Poisson prediction
x=5
p=.25
y
10
20
30
40
50

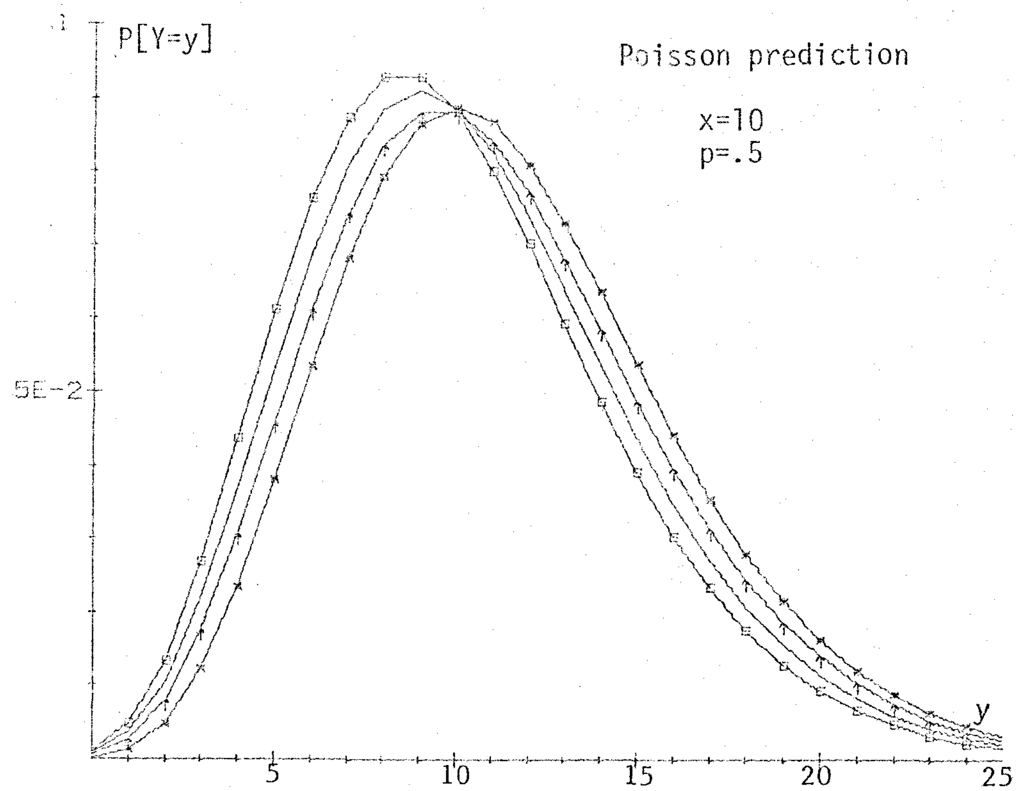

P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Poisson prediction
x=5
p=.25
y
10
20
30
40
50


P[Y=y]
Poisson prediction
x=10
p=.25
3E-2
2E-2
1E-2
y
25
50
75

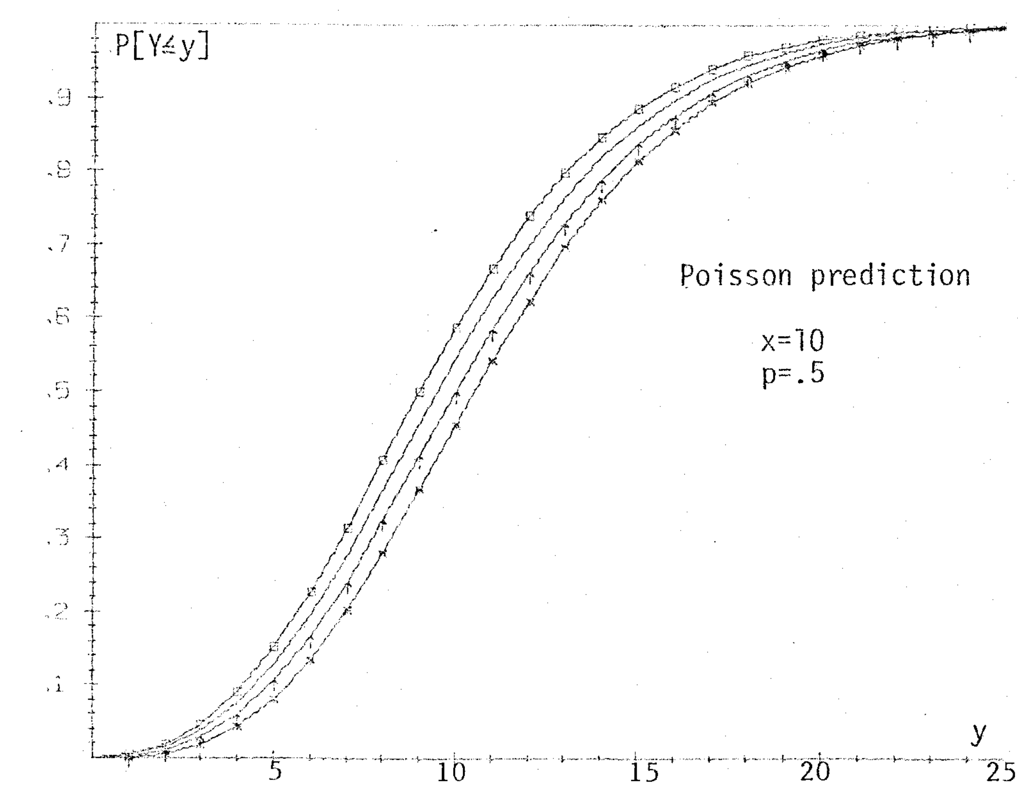

P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Poisson prediction
x=10
p=.25
y
25
50
75


.15
P[Y=y]
Poisson prediction
x=5
p=.5
.1
5E-2
y
5
10
15
20

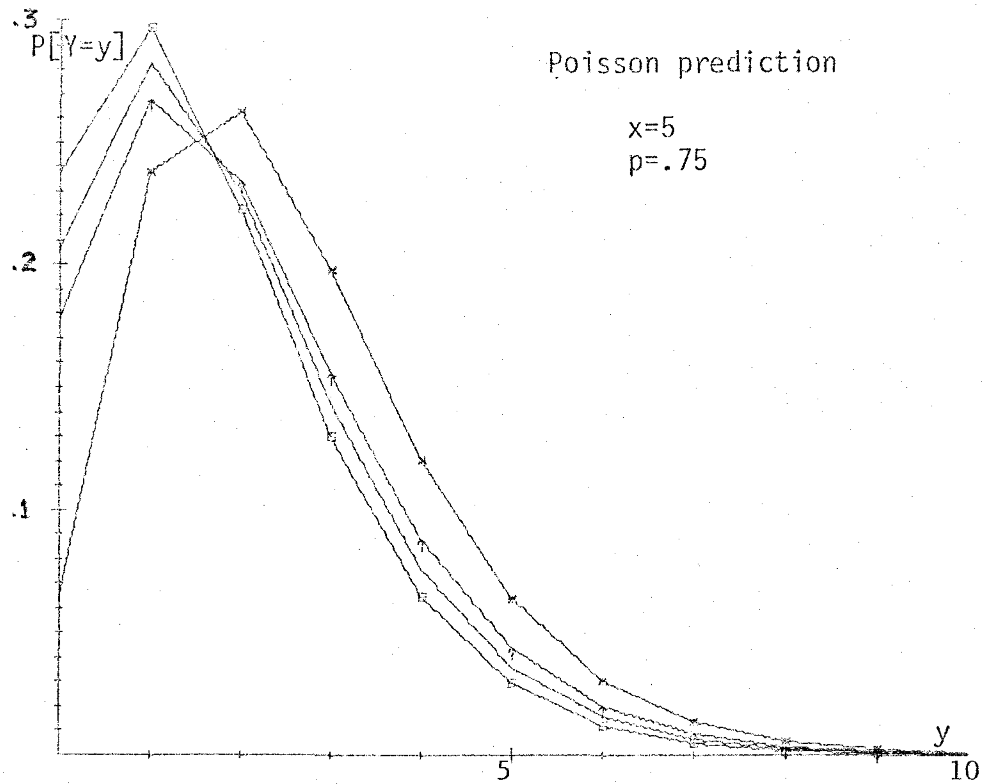

P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Poisson prediction
x=5
p=.5
y
5
10
15
20

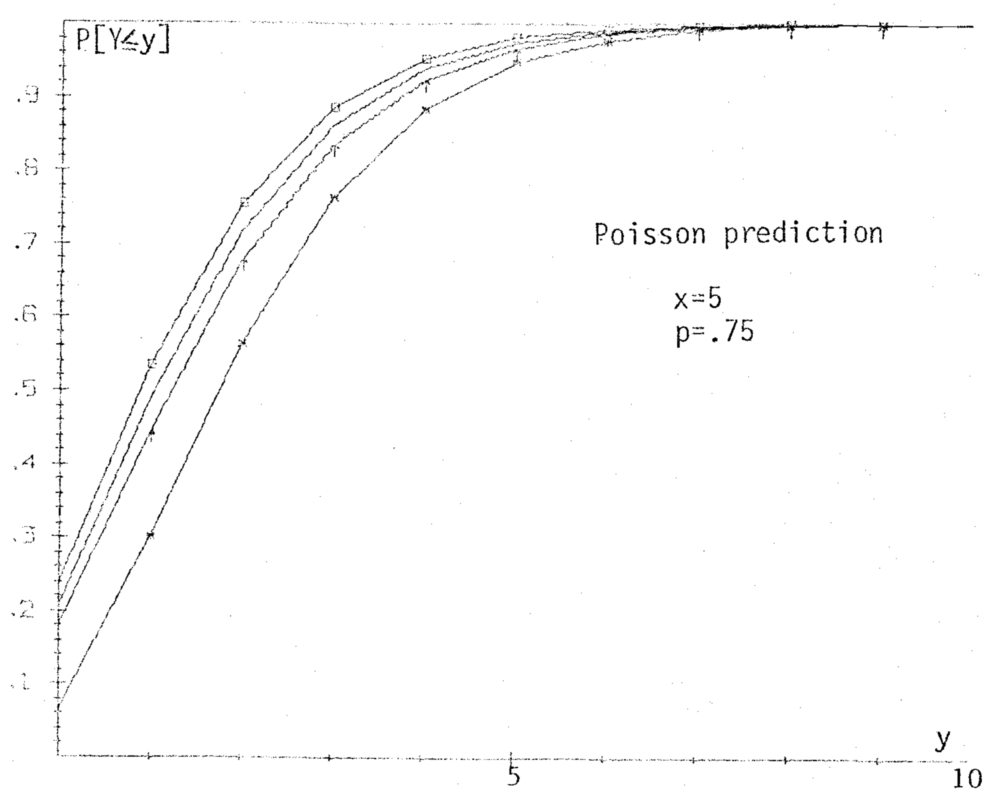


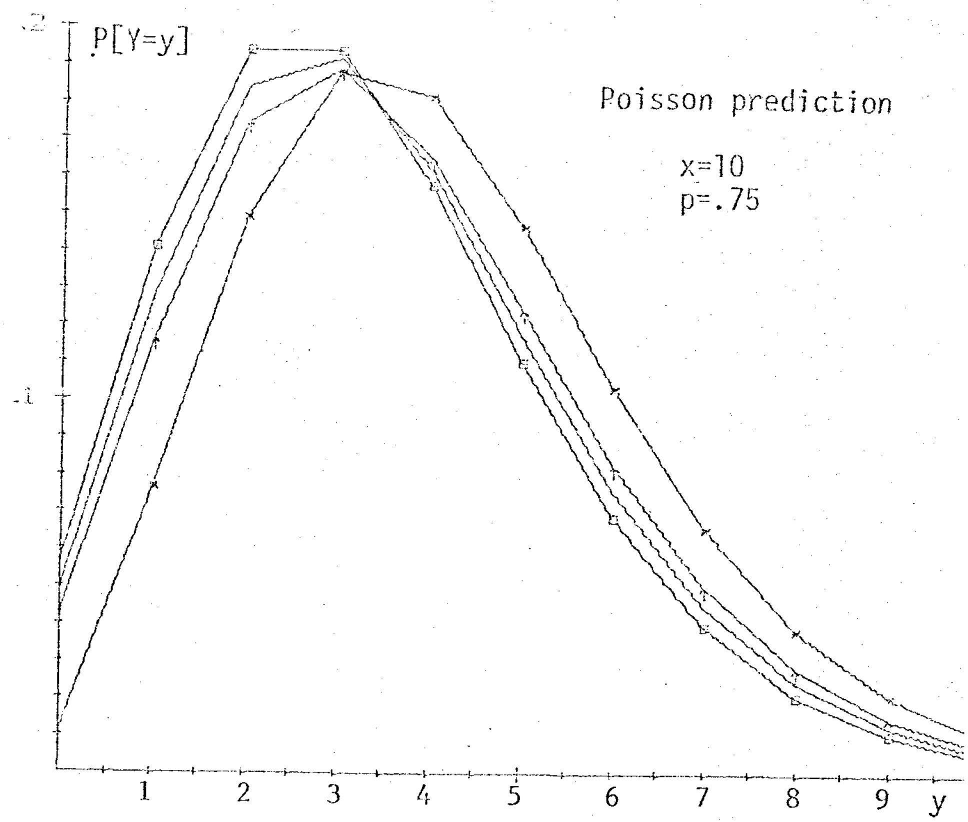
P[Y=y]
Poisson prediction
x=10
p=.5
.1
5E-2
y
5
10
15
20
25

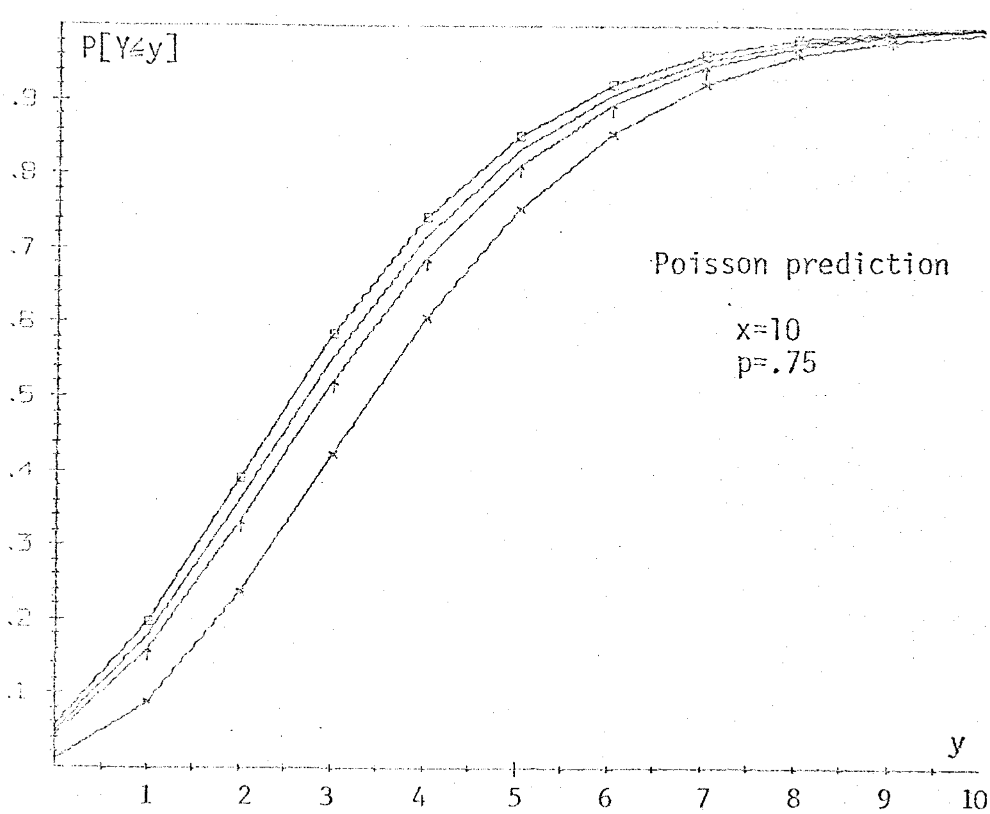


P[Y≤y]
Poisson prediction
x=10
p=.5
.9
.8
.7
.6
.5
.4
.3
.2
.1
y
5
10
15
20
25


.3
P[Y=y]
Poisson prediction
x=5
p=.75
.2
.1
y
5
10


P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Poisson prediction
x=5
p=.75
y
5
10


P[Y=y]
Poisson prediction
x=10
p=.75
.2
.1
1
2
3
4
5
6
7
8
9
y


P[Y≤y]
Poisson prediction
x=10
p=.75
.9
.8
.7
.6
.5
.4
.3
.2
.1
y
1
2
3
4
5
6
7
8
9
10

III.8 Conclusions

Because it falls between a conservative and an anticonservative predictive distribution the MLPD is expected to yield prediction intervals whose exact probabilities of coverage $\beta(\lambda)$ are close to the nominal level of confidence for any given λ. One could in fact compute $\beta(\lambda)$ and compare it to the nominal level; this has been done for the binomial case (see section V.7).

Since the MLPD is approximately a $NB[x + \frac{1}{2}, n/(m + n)]$ distribution it can be said to lie between the conservative and the anticonservative distributions, which are $NB(x, n/(n + m))$ and $NB(x + 1, n/(n + m))$. It is noticeable that the three corresponding Bayes priors are the priors most frequently encountered in the literature around the classical theories, i.e., the λ^{-1}, $\lambda^{-\frac{1}{2}}$ and uniform priors.

The asymptotic results that have been established guarantee for the three distributions the convergence towards the true distribution of the sum Y, for n tending to infinity. As for Fisher's approach it tends very rapidly towards the $NB(X + 1, p)$ distribution when mx/n (i.e., roughly the expectation of Y) increases. All approaches tend to be equivalent to the usual normal approximation approach when both n and $m \to \infty$.

Another important result is that the predictions depend on n and m only through their ratio or equivalently through $p = n/(n + m)$. The larger the value of p the closer will be the MLPD, the $NB(x,p)$ and the $NB(x + 1, p)$ distributions. Note that in the finite population framework described in section III.1 , p is the fraction of the of the population that has been sampled.

IV STRATIFIED RANDOM SAMPLING FROM A POISSON POPULATION

In this chapter the methods exposed earlier are applied to the prediction of the grand sum of future random samples drawn from Poisson strata. When the ratio of the observed sample size to the future sample size is constant over the strata the problem is the same as in chapter III. The results given for the MLPD approach illustrate how helpful this approach might become in some rather complex situations.

IV.1 Problem and notations

Suppose we have a population composed of k strata, each stratum having a Poisson distribution with parameter λ_i. We sample at random from each stratum; n_i and x_i respectively denote the sample size and the sample sum for stratum i. We are interested in making predictions on the grand total Y of a future sample of size $\sum_{i=1}^{k} m_i$, where m_i elements are to be sampled from stratum i.

We introduce further the following notations:

$$X = \sum_{i=1}^{k} X_i$$

$$Y = \sum_{i=1}^{k} Y_i$$

where Y_i is the sum of the "future" sample from stratum i,

$$T = X + Y$$

$$p_i = n_i/(m_i + n_i) \qquad q_i = 1 - p_i \quad .$$

It is to be noted that the Y_i's and the X_i's are sufficient statistics for $\underset{\sim}{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)'$.

IV.2 Frequentist approach

Let us consider the joint distribution of $(Y, X_1, X_2, \ldots X_k)$

$$f(y, x_1, x_2, \ldots, x_k) = \frac{1}{y!}\, e^{\sum_{i=1}^{k}(m_i + n_i)\lambda_i} \left(\sum_{i=1}^{k} m_i\lambda_i\right)^y \prod_{i=1}^{k} \frac{(n_i\lambda_i)^{x_i}}{x_i!}$$

In the general case the frequentist conditional method fails to apply because there is no tractable function $\psi(y, x_1, x_2, \ldots, x_k)$ such that conditioning Y on ψ leads to the elimination of the λ_i's . However in the following special case where the m_i's are proportional to the n_i's , there is an easy solution.

Special case: p_i = constant p

We investigate the joint distribution of X and Y . We have

$$X \sim P(\beta) \qquad \text{where} \qquad \beta = \sum_{i=1}^{k} n_i\lambda_i$$

$$Y \sim P(\alpha) \qquad \text{where} \qquad \alpha = \sum_{i=1}^{k} m_i\lambda_i$$

$$f_{X,Y}(x,y) = \frac{e^{-\beta}\beta^x}{x!} \; \frac{e^{-\alpha}\alpha^y}{y!}$$

Thus for $T = X + Y$ we have

$$f_{Y|T} = \binom{t}{y} \frac{\beta^{t-y}\alpha^y}{(\alpha + \beta)^t} \quad .$$

We see that in order to get for $Y|T$ a distribution independent of the λ_i's we need

$$\frac{\alpha}{\alpha + \beta} = \frac{\sum_{i=1}^{k} m_i \lambda_i}{\sum_{i=1}^{k} (m_i + n_i)\lambda_i} \qquad \text{independent of} \quad \lambda_i .$$

That is, $m_i/(m_i + n_i) = q$ constant. Then

$$Y|Y + X \sim B(Y + X, q) ,$$

which in turn means that when the ratio of the size of the observed sample to the size of the future sample is constant for each stratum, the problem can be viewed as simple random sampling as far as the frequentist approach is concerned.

Consequently the results established in section III.2 are applicable here and in particular the prediction limits are to be read from a $NB(x,p)$ for the lower limit and a $NB(x + 1,p)$ for the upper limit.

IV.3 Bayesian approach

IV.3.1 Preliminary remark

Suppose X and Y are discrete random vectors with mass probability functions $f_X(x|\theta)$ and $f_Y(y|\theta)$ respectively, θ being a common k-dimensional unknown parameter belonging to Θ.

Let us observe an outcome x of X and derive from it, by a Bayesian argument, a predictive distribution for a function $Z = \psi(Y)$.

Let us denote by $\pi(\theta|x)$ the posterior density of θ. Then

$$f_{Z|X}(z|x) = \int_{\Theta} f_Z(z|\theta)\,\pi(\theta|x)d\theta$$

$$f_{Z|X}(z|x) = \int_{\Theta} \{\sum_{\mathcal{Y}_z} f_Y(y|\theta)\}\,\pi(\theta)d\theta \tag{4.1}$$

with $\mathcal{Y}_z = \{y: \psi(y) = z\}$

But since we deal with probabilities the properties of absolute convergences required to interchange the summation and integral signs are verified so that equivalent to (4.1) we have

$$f_{Z|X}(z|x) = \sum_{\mathcal{Y}_z} \int_{\Theta} f_Y(y|\theta)\,\pi(\theta|x)d\theta \tag{4.2}$$

Applying this to the stratified sampling problem it implies that we are allowed to get the predictive distribution for each Y_i separately first, and then by means of convolutions to derive the predictive distribution for their sum Y. This will be much easier than going the direct way according to (4.1).

IV.3.2 Uniform and $1/\lambda_i$ priors.

Each stratum being regarded as a population we obtain

$Y_i|X_i \sim NB(x_i + 1;p_i)$ for a uniform prior,

$Y_i|X_i \sim NB(x_i;p_i)$ for a $1/\lambda_i$ prior .

But the distribution of the sum of random variables having negative binomial distributions with different parameters p_i does not take a familiar form. Thus we have to rely on a computational device to produce the results of the convolutions. For example, with two strata if we denote

$$\Pr\{Y_1 = i | x_1\} = a_i$$

and

$$\Pr\{Y_2 = j | x_2\} = b_j \quad ,$$

we compute for each value of s

$$\Pr\{Y = s | x_1, x_2\} = \sum_{i=0}^{s} a_i b_{s-i}$$

Such computations are reasonable for a small number of strata and a small value for $\sum_{i=1}^{k} (m_i x_i / n_i)$ which indicates the central location of the distribution of Y , but they become rapidly prohibitive when these increase.

IV.3.3 Special case p_i = constant p

Suppose $Z_i \sim NB(_i, \theta)$, $i = 1,2,\ldots,k$. Then

$$\sum_{i=1}^{k} Z_i \sim NB(\sum_{i=1}^{k} r_i ; \theta) .$$

This is easily seen from the moment generating functions $\left(\frac{\theta}{1 - (1 -)e^t}\right)^{r_i}$.

Consequently in this special case we have

$Y|x \sim NB(x + k,p)$ for uniform priors,

$Y|x \sim NB(x,p)$ for $1/\lambda_i$ priors, for all i,

and

$Y|x \sim NB(x + s,p)$ when choosing a uniform

prior for any s or the λ_i's and a $1/\lambda_i$ prior for the $k - s$ remaining λ_i's.

Thus the lower prediction limit in the frequentist approach is the same as for the Bayesian approach with priors $1/\lambda_i$ for all i and the upper frequentist limit is the same as for the Bayesian approach with priors $1/\lambda_i$ for all but one stratum for which a uniform prior is chosen.

IV.4 Likelihood prediction

IV.4.1 General formulation

We have to maximize over $\underset{\sim}{\lambda}(\lambda_i,\ldots,\lambda_k)'$ the quantity

$$f_Y(y;\textstyle\sum m_i\lambda_i)f_{\underset{\sim}{X}}(\underset{\sim}{x};\underset{\sim}{\lambda}) = \frac{1}{y!}e^{-\sum m_i\lambda_i}(\textstyle\sum m_i\lambda_i)^y \prod_{i=1}^{k} \frac{e^{-n_i\lambda_i}(n_i\lambda_i)^{x_i}}{x_i!}$$

Taking the logarithms, we obtain for this expression

$$\sum_{i=1}^{k}(-n_i\lambda_i + x_i\log(n_i\lambda_i) - \log x_i! - m_i\lambda_i) + y\log(\textstyle\sum m_i\lambda_i) - \log y!\ ,$$

and taking the derivative with respect to λ_i we have

$$-n_i + \frac{x_i}{\lambda_i} - m_i + y\,\frac{m_i}{\sum m_j\lambda_j} = 0 \qquad i = 1,\ldots,k\ .$$

Thus the maximum likelihood estimators $\hat{\lambda}_i$, $(i = 1,2,\ldots,k)$ satisfy

$$x_i - n_i\hat{\lambda}_i = m_i\hat{\lambda}_i(1 - y/\sum_{j=1}^{k} m_j\hat{\lambda}_j) \qquad i = 1,\ldots,k$$

By summing over i we get the relationship:

$$x + y = \sum_{i=1}^{k}(n_i + m_i)\hat{\lambda}_i \quad , \tag{4.3}$$

and we have

$$f_Y(y;\sum m_i\hat{\lambda}_i)f_{\underset{\sim}{X}}(\underset{\sim}{x};\underset{\sim}{\hat{\lambda}}) = \frac{1}{y!}\ e^{-(x+y)}(\sum_{i=1} m_i\hat{\lambda}_i)^y \prod_{i=1}^{k} \frac{(n_i\hat{\lambda}_i)^{x_i}}{x_i!}$$

Finally the prediction likelihood of Y is obtained by solving the system

$$\hat{L}(y;x) \propto \frac{e^{-y}(\sum_{i=1}^{k} m_i\hat{\lambda}_i)^y}{\Gamma(y+1)} \prod_{i=1}^{k} (\hat{\lambda}_i)^{x_i} \tag{4.4a}$$

$$x_i - n_i\hat{\lambda}_i = m_i\hat{\lambda}_i(1 - y/\sum_{j=1}^{k} m_j\hat{\lambda}_j) \tag{4.4b}$$

IV.4.2 Existence of a MLPD.

Theorem 4.1: The maximum likelihood prediction method applied to stratified random sampling from a Poisson population always allows to define a Maximum Likelihood Prediction Distribution.

Proof: We want to show that $\sum_{y=0}^{k} \hat{L}(y;\underset{\sim}{x}) < \infty$ for all $x_i \in I = \{0,1,2,\ldots\}$, with

$$\hat{L}(y;\underset{\sim}{x}) = c(\underset{\sim}{x}) \frac{1}{y!} e^{-y}\left(\sum_{i=1}^{k} m_i\lambda_i\right)^y \prod_{i=1}^{k} (\hat{\lambda}_i)^{x_i}$$

From (4.3) we have

$$\sum_{i=1}^{k} m_i\hat{\lambda}_i(1 + n_i/m_i) = x + y$$

Let $\varphi = \mathrm{Min}[n_i/m_i]$, then

$$(1 + \varphi) \sum_{i=1}^{k} m_i\hat{\lambda}_i \le x + y$$

$$\sum_{i=1}^{k} m_i\hat{\lambda}_i \le \theta(x + y) \quad \text{where} \quad \theta = (1 + \varphi)^{-1}, \ 0 < \theta < 1 , \tag{4.5}$$

and also $\hat{\lambda}_i \le x + y$. (4.6)

Using (4.5) and (4.6) we obtain

$$\hat{L}(y;\underset{\sim}{x}) \le c_1(\underset{\sim}{x}) \frac{1}{y!} e^{-y}[\theta(x + y)]^y \prod_{i=1}^{k} (x + y)^{x_i}$$

$$\hat{L}(y;\underset{\sim}{x}) \le c_1(\underset{\sim}{x}) \frac{1}{y!} e^{-y} (x + y)^{x + y}\theta^y$$

The function of y on the right side decreases faster than a $NB(x+1,\theta)$ density so that by theorem 3.4 there exists a MLPD.

IV.4.3 A proposed general solution.

In order to determine the MLPD, $\hat{F}(y;x)$, we need to solve the system (4.4a),(4.4b) where we simply disregard the normalization constant for (4.4a) .

Let us set $k = 1 - y/\sum_{j=1} m_j \hat{\lambda}_j$. Then (4.4b) gives

$$\hat{\lambda}_i = \frac{x_i}{n_i + m_i k}$$

We note that k can range from $-\mathrm{Min}[n_i/m_i]$ to 1 , the corresponding variation for y being as illustrated below.

k	$-\mathrm{Min}(n_i/m_i)$	0	1
y	$+\infty$	$\sum_{i=1}^{k}(m_i x_i/n_i)$	0

Thus we solve (4.4a),(4.4b) by a two stage procedure:

i) pick a value k in the interval $[-\mathrm{Min}(n_i/m_i), 1]$

ii) compute $\hat{\lambda}_i = x_i/(n_i + m_i k)$ for $i = 1,2,\ldots,k$

iii) compute $y = (1 - k)\sum_{i=1}^{k} m_i \hat{\lambda}_i$

iv) compute $\hat{L}(y;\underset{\sim}{x})$ from (4.4a) .

The main disadvantage of the procedure is that the likelihood is evaluated at arbitrary values of y. We need to interpolate for the integer values in order to derive the probability mass function.

Nevertheless contrary to the computations involved for the Bayesian prediction this procedure does not become more tedious when the number of strata increases.

IV.4.4 Special case: p_i = constant p

From (4.3) we obtain

$$\sum_{j=1}^{k}(n_j + m_j)\hat{\lambda}_j = x + y$$

$$(1 + p/q) \sum_{j=1} m_j\hat{\lambda}_j = x + y$$

$$\sum_{j=1}^{k} m_j\hat{\lambda}_j = q(x + y) \quad ,$$

so that we are able now to solve (4.4b)

$$\hat{\lambda}_i \left[m_i + n_i - \frac{m_i y}{q(x + y)}\right] = x_i$$

$$\hat{\lambda}_i(n_i + m_i)(1 - \frac{y}{x + y}) = x_i$$

$$\hat{\lambda}_i = \frac{x_i}{n_i + m_i} \; \frac{x + y}{x} \quad .$$

Finally, substituting in (4.4a) we have

$$\hat{f}(y;x) \propto \frac{e^{-y} (x + y)^{x + y} q^y}{y!} \quad ,$$

which is the same result as for simple random sampling.

IV.5 Conclusions

In the case where the sampling ratios (n_i/m_i) are the same for all strata the same conclusions as in simple random sampling can be drawn.

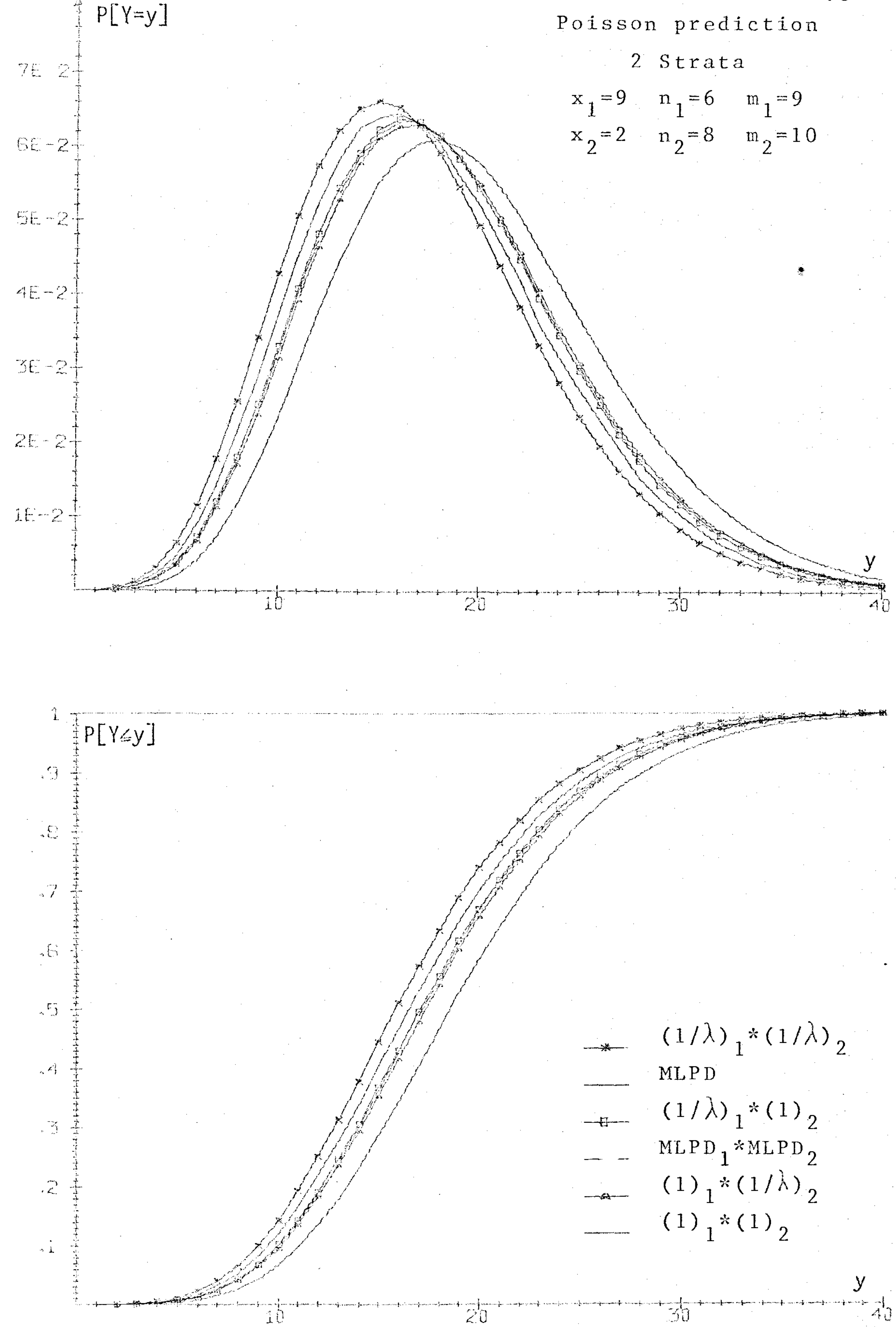
However for the Bayesian prediction we then have to chose a uniform prior for any one of the λ_i's and $1/\lambda_i$ priors for the $(k - 1)$ remaining strata in order to obtain the upper frequentist limits. In fact it would be equivalent but less convenient to choose a $(1/\lambda_i)^{(k-1)/k}$ prior for each stratum. This becomes necessary when we extend the Bayesian prediction to the general case where n_i/m_i are no longer constant since it matters then which λ_i is chosen to be given the uniform prior, and the $(1/\lambda_i)^{(k-1)/k}$ prior avoids having to discriminate among strata.

In this general case we expect the MLPD will still be located between the two Bayesian distributions $1/\lambda_i$ and $(1/\lambda_i)^{(k-1)/k}$ for all i because of the smooth variation of these functions with respect to n_i/m_i, as illustrated by the example below.

We note that for $k = 2$ the $(1/\lambda_i)^{(k-1)/k}$ priors are $\lambda_i^{-\frac{1}{2}}$ priors and the corresponding Bayesian prediction is very close to the prediction obtained by convoluting the individual MLPD's of each stratum, a method that will be considered in the binomial case.

As far as the computations are concerned for the general case they will be tedious and time-consuming for the Bayesian approach with more than 2 or 3 strata and a large $\sum_{i=1}^{k} m_i x_i/n_i$. For the MLPD the computations as proposed in IV.4.3 are not as simple but do not increase with the number of strata and increase little with larger $\sum_{i=1}^{k} (m_i x_i/n_i)$.

The following example exhibits for 2 strata the predictions for the 4 possible combinations of uniform and $1/\lambda$ priors, for the MLPD and for the distribution resulting from the convolution of the MLPD of each stratum, i.e., approximately the Bayesian prediction for priors $(1/\lambda_i)^{-\frac{1}{2}}$, $i = 1,2\ldots$.


P[Y=y]
Poisson prediction
2 Strata
x1=9 n1=6 m1=9
x2=2 n2=8 m2=10
7E-2
6E-2
5E-2
4E-2
3E-2
2E-2
1E-2
y
10
20
30
40
P[Y≤y]
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
(1/λ)1*(1/λ)2
MLPD
(1/λ)1*(1)2
MLPD1*MLPD2
(1)1*(1/λ)2
(1)1*(1)2

V. SIMPLE RANDOM SAMPLING FROM A BINOMIAL POPULATION

The developments for the binomial population follow the same pattern as these in chapter III concerning the Poisson population.

The MLPD compares to the frequentist conditional and the Bayesian approaches in very much the same way. The study of asymptotic behaviors leads also to similar results. More studies have been devoted to the binomial problem and we will refer mainly to the works of Fisher (1959), Thatcher (1964) and Olsen (1974).

V.1 Problem

We assume that we have an infinite (or finite, see section III.1) population in which each unit presents a characteristic A (referred to as "success") with probability p and presents not-A with probability q .

We sample n units at random from this population and observe x successes among them. What is then the probability of observing Y successes in m future trials?

The number of successes can be readily viewed as a sample sum by introducing for each trial the Bernoulli random variable.

V.2 Frequentist approach

Olsen (1974) shows that, x and Y being the variables defined above, one can write

$$\Pr\{Y \le k | T = x + k\} = \Pr\{Z \le k\} \tag{5.1}$$

where Z has a $NH(x,n,m)$ distribution as defined below.

Definition 5.1: A random variable Z is said to have a Negative Hypergeometric distribution with parameters (x,n,m), noted $NH(x,n,m)$, when its probability mass function is:

$$f_Z(z) = \frac{\binom{n}{x-1}\binom{m}{z}(n-x+1)}{\binom{m+n}{z+y-1}(m+n-x-z+1)} \qquad \text{for } z = 0,1,\ldots,m \quad .$$

Some properties of this distribution are given in Section V.6.2. Applying to (5.1) corollaries (2.3) and (2.5) the next theorem follows immediately.

Theorem 5.2: The α-upper prediction limit given by the frequentist conditional approach for the number of successes among m items sampled at random from a binomial population, given x successes out of n items, is the α-upper fractile of the $NH(x+1,n,m)$ distribution (except for $x = n$). The β-lower prediction limit is the β-fractile of the $NH(x,n,m)$ distribution (except for $x = 0$) .

V.3 Bayesian approach.

We show now the result corresponding to theorem (3.2) of Section III.3 for the Poisson case.

Theorem 5.3: The frequentist conditional lower and upper prediction limits coincide respectively with the Bayes lower prediction limit under $1/(1-p)$ prior.

Proof: We have

$$f_X(x;p)f_Y(y;p) = \binom{n}{x}\binom{m}{y} p^{x+y}(1-p)^{m+n-x-y} .$$

For a Bayesian $1/p$ prior we get, assuming $x \geq 1$,

$$f_{1/p}(y|x) = \frac{\binom{n}{x}\binom{m}{y} \int_0^1 p^{x+y-1}(1-p)^{m+n-x-y}dp}{\binom{n}{x} \int_0^1 p^{x-1}(1-p)^{n-x}dp}$$

$$f_{1/p}(y|x) = \frac{\binom{m}{y} \ (x+y-1)! \ (m+n-x-y)!/(m+n)!}{(x-1)! \ (n-x)! \ /n!}$$

which is the probability mass function of a NH(x,n,m) distribution. For a $1/(1-p)$ prior we have, assuming $x \leq n-1$,

$$f_{1/1-p}(y|x) = \frac{\binom{n}{x}\binom{m}{y} \int_0^1 p^{x+y}(1-p)^{m+n-x-y-1}dp}{\binom{n}{x} \int_0^1 p^{x}(1-p)^{n-x-1}dp}$$

$$f_{1/1-p}(y\ x) = \frac{\binom{m}{y} \ (x+y)! \ (m+n-x-y-1)! \ /(m+n)!}{x! \ (n-x-1)! \ / \ n!}$$

which is a NH(x + 1,n,m) distribution.

These priors are the same as the priors used to obtain the usual upper and lower confidence limits for the parameter of a binomial distribution in the classical theory of confidence intervals (see Pratt (1965).) Thatcher (1964) showed that there is no prior that gives limits coinciding with the frequentist limits.

In the following we denote $f_{1/p}(y|x)$ by $f_1(y|x)$ and $f_{1/1-p}(y|x)$ by $f_2(y|x)$ and correspondingly $F_1(y|x)$ and $F_2(y|x)$ for the CDF's.

More generally, for a $\text{BETA}(\alpha,\beta)$ prior, i.e.

$$\pi(p) \propto p^{\alpha - 1}(1 - p)^{\beta - 1}$$

we have

$$f(y|x) \propto \binom{m}{y} \int_0^1 p^{x + y + \alpha - 1}(1 - p)^{m + n - x - y + \beta - 1} dp$$

$$\propto \binom{m}{y} \Gamma(x + y + \alpha)\, \Gamma(m + n - x - y + \beta)$$

which, by extending the definition of section V.2 to non-integer values of the first and second parameter, is a $NH(x+\alpha, n+\alpha+\beta-1, m)$ distribution. This is a special class of the Beta-Binomial family where the parameter p of a $B(m,p)$ distribution is taken to have a $\text{BETA}(x + \alpha,\ n - x + \beta)$ distribution.

V.4 Likelihood approaches.

V.4.1 Fisher's likelihood.

The expression given by Fisher (1959) is

$$\hat{L}(y;x) = \frac{n^n m^m}{N^N} \frac{(x + y)^{x + y}(N - x - y)^{N - x - y}}{x^x y^y (n - x)^{n - x}(m - y)^{m - y}}$$

where $N = m + n$.

The maximum, 1, is attained as in the Poisson case for $\tilde{y} = mx/n$.

V.4.2 Prediction Likelihood approach

The MLE of p based on x and y is $\hat{p}_{xy} = (x + y)/(m + n)$, so that

$$f_X(x;\hat{p}_{xy})f_y(y;\hat{p}_{xy}) = \binom{n}{x}\binom{m}{y}\frac{x+y}{m+n}^{x+y}(1 - \frac{x+y}{m+n})^{m+n-x-y}$$

$$= N^{-N}\binom{n}{x}\binom{m}{y}(x+y)^{x+y}(N-x-y)^{N-x-y} ,$$

or $$\hat{L}(y;x) \propto \binom{m}{y}(x+y)^{x+y}(N-x-y)^{N-x-y} .$$

Since y takes only a finite number of values, from 0 to m , it is always possible to define a Maximum Likelihood Prediction Distribution (MLPD).

V.4.3 Comparison of the two Likelihoods

Comparing the ratio of Fisher's expression to the Prediction Likelihood Function, as a function of y we have

$$\tilde{L}(y;x)/\hat{L}(y;x) \propto \frac{y!\ (m-y)!}{y^y\ (m-y)^{m-y}} ,$$

or, approximately, using Stirling's formula,

$$\tilde{L}(y;x)/\hat{L}(y;x) \quad \alpha \quad [y(m-y)]^{\frac{1}{2}} .$$

The ratio is no longer monotone as in the Poisson case, so that nothing can be said about the relative locations of the two likelihoods.

By taking the logarithm of $\hat{L}(y;x)$ and its derivative with respect to y , it is easily established that the maximum of the prediction likelihood is attained for $\hat{y}$ such that

$$\frac{(x + \hat{y})(m - \hat{y})}{(N - x - \hat{y})\,\hat{y}} = \exp\{1/\hat{y} - 1/(m - \hat{y})\} \quad .$$

For large values of $\hat{y}$ and $m - \hat{y}$ the right hand side is approximately equal to one and $\hat{y} \cong mx/n$. That is to say that the two maxima $\hat{y}$ and $\tilde{y}$ will tend to be the same only when $\hat{y}$ and $(m - \hat{y})$ are large which occurs when m and n are large.

V.5 Comparison between the MLPD and the Bayesian predictions

In order to establish the main theorem of this section we need the following lemma.

Lemma 5.4: The finite sequences $\{u_t;\ t = 0,1,\ldots,N - 1\}$ and $\{v_t;\ t = 1,2,\ldots,N\}$ are increasing, where

$$u_t = \frac{t!(N - t - 1)!}{t^t(N - t)^{N - t}}$$

and

$$v_t = \frac{t^t(N - t)^{N - t}}{(t - 1)!(N - t)!}$$

Proof: First note that since $v_t = 1/u_{N - t}$, it suffices to show that u_t is increasing. Let us look at the difference

$$\log u_{t+1} - \log u_t = \log(t+1) - \log(N-t-1) - (t+1)\log(t+1) + t\log t$$

$$- (N-t-1)\log(N-t-1) + (N-t)\log(N-t)$$

$$= (N-t)\log\frac{N-t}{N-t-1} - t\log\frac{t+1}{t} \tag{5.2}$$

$$= (N-t)\log\left(1+\frac{1}{N-t-1}\right) - t\log\left(1+\frac{1}{t}\right),\ t = 0,1,2\ldots N\text{-}2.$$

But since $\log(1+x) \leq x$ for all x, we have

$$t\log(1+1/t) \leq 1 \quad \text{for} \quad t = 0,1,\ldots,N-2 \ .$$

Also we have seen in Section II.5 that for $x > 0$,

$$(x+1)\log(1+1/x) > 1 \ ,$$

so that

$$(N-t)\log\left(1+\frac{1}{N-t-1}\right) > 1 \quad \text{for} \quad t = 0,1,\ldots,N-2 \ .$$

Thus from (5.2) the desired result follows:

$$\log u_{t+1} - \log u_t > 0 \quad \text{for} \quad t = 0,1,\ldots,N-2 \ .$$

Theorem 5.5: When sampling from a Binomial distribution the following inequalities hold between the CDF's of the MLPD and the Bayesian predictive distributions with priors $1/p$ and $1/(1-p)$:

$$F_2(y|x) \leq \hat{F}(y;x) \leq \hat{F}_1(y|x)$$

for all values of x,m,n for which F_1 and F_2 exist.

Proof: Consider first the ratio:

$$\frac{f_2(y|x)}{\hat{f}(y;x)} \propto \frac{(x + y)!\ (N - x - y - 1)!}{(x + y)^{x + y}(N - x - Y)^{N - x - y}}$$

with $0 \leq x \leq n - 1$ and $0 \leq y \leq m$. Setting $t = x + y$, we have $0 \leq t \leq N - 1$, and the ratio is proportional to u_t and thus is an increasing function of y for any fixed x. This proves the left inequality by Theorem 3.4 . We now turn to

$$\frac{\hat{f}(y;x)}{f_1(y|x)} \propto \frac{(x + y)^{x + y}(N - x - y)^{N - x - y}}{(x + y - 1)!\ (N - x - y)!}$$

with $0 \leq x \leq n - 1$ and $0 \leq y \leq m$.

By recognizing for the right hand side the sequence v_t , where $t = x + y$, we see that the ratio of probability mass functions is increasing and thus the right inequality holds.

Existence of a Bayes prior yielding the MLPD

The problem of finding a Bayes prior yielding $\hat{f}(y;x)$ reduces to finding a non-negative function g such that

$$\int_0^1 p^t(1 - p)^{N - t}g(p,N)dp = k(N)t^t(N - t)^{N - t} ,$$

or equivalently, by setting $p/(1 - p) = e^{-u}$, to finding a positive function $h(u,N)$ such that

$$\int_{-\infty}^{+\infty} e^{-ut}h(u,N)du = t^t(N - t)^{N - t} .$$

As in the Poisson case it is difficult to obtain the exact solution and we work out an approximation.

Recalling from (3.3) that:

$$\Gamma(t + \tfrac{1}{2}) = (2\pi)^{\frac{1}{2}} e^{-t} t^{t} \exp\{-\frac{1}{24t} + \frac{7}{8(360t^3)} + o(\frac{1}{t^5})\} \quad (5.3)$$

we derive the following relationship:

$$\Gamma(t + \tfrac{1}{2})\Gamma(N - t + \tfrac{1}{2}) = 2\pi\, e^{-N} t^{t} (N - t)^{N - t} \exp\{-\frac{1}{24t} - \frac{1}{24(N-t)} + \ldots\} \quad (5.4)$$

i.e., approximately

$$\Gamma(t + \tfrac{1}{2})\Gamma(N - t + \tfrac{1}{2}) \cong 2\pi\, e^{-N} t^{t} (N - t)^{N - t} (1 - \frac{1}{24t} - \frac{1}{24(N-t)}) \quad (5.5)$$

so that prior $[p(1 - p)]^{\frac{1}{2}}$ comes fairly close to $\hat{f}(y;x)$ since

$$\int_0^1 p^t (1 - p)^{N - t} [p(1-p)]^{-\frac{1}{2}} = \Gamma(t + \tfrac{1}{2})\Gamma(N - t + \tfrac{1}{2}) \,/\, \Gamma(n + 1)$$

$$\cong 2\pi \frac{e^{-N}}{N!} t^{t} (N - t)^{t} \exp\{-\frac{1}{24t} - \frac{1}{24(N-t)}\} .$$

As was seen in Section V.3 this prior yields a $NH(x + \frac{1}{2}, n, m)$ predictive distribution.

<u>Determination of the normalization constant for $\hat{f}(y;x)$</u>

We first establish the expression of the probability mass function of a $NH(x + \frac{1}{2}, n, m)$ distribution. For a prior $\pi_0(p) = [p(1-p)]^{-\frac{1}{2}}$ we have

$$f_{\pi_o}(y|x) = \frac{\int_o^1 \binom{n}{x}\binom{m}{y} p^{x+y-\frac{1}{2}}(1-p)^{N-x-y-\frac{1}{2}}dp}{\int_o^1 \binom{n}{x} p^{x-\frac{1}{2}}(1-p)^{n-x-\frac{1}{2}}dp}$$

$$\frac{n!\binom{m}{y}\Gamma(x+y+\frac{1}{2})\,\Gamma(N-x-y+\frac{1}{2})}{N!\,\Gamma(x+\frac{1}{2})\Gamma(n-x+\frac{1}{2})}$$

Thus

$$\frac{n!}{N!\Gamma(x+\frac{1}{2})\Gamma(n-x+\frac{1}{2})} \sum_{y=0}^{m} \binom{m}{y} \Gamma(x+y+\frac{1}{2})\Gamma(N-x-y+\frac{1}{2}) = 1 .$$

Now using approximation (5.5) we obtain

$$\frac{2\pi e^{-N} n!}{N!\ \Gamma(x+\frac{1}{2})\Gamma(n-x+\frac{1}{2})} \sum_{y=0}^{m} \binom{m}{y} (x+y)^{x+y}(N-x-y)^{N-x-y}$$

$$* \ (1 - \frac{1}{24(x+y)} - \frac{1}{24(N-x-y})\ ) \doteq 1 ,$$

or approximately

$$\frac{2\pi e^{-N} n!}{N!\Gamma(x+\frac{1}{2})\Gamma(n-x+\frac{1}{2})} \sum_{y=0}^{m} \binom{m}{y} (x+y)^{x+y}(N-x-y)^{N-x-y}$$

$$\cong 1 + \frac{1}{24} E[\frac{1}{x+y} + \frac{1}{N-x-y}] \qquad (5.6)$$

where the expectation is the expectation of the given function of y with respect to the MLPD. We work out for it the following approximations. First we compute $E[\frac{1}{x+y}]$ for y having a $NH(x+1,n,m)$ distribution.

$$E\left[\frac{1}{x+y}\right] = \sum_{y=0}^{m} \frac{\binom{m}{y} \quad (x+y-1)!\,(m+n-x-y-1)!/(m+n)!}{x!\ (n-x-1)!\ /\ n!}$$

$$= \frac{n}{x(m+n)} \sum_{y=0}^{m} \frac{\binom{m}{y} \quad (x+y-1)!\,(m+n-1-x-y)!/(m+n-1)!}{(x-1)!\,(n-1-x)!/(n-1)!}$$

The right hand side is the summation of the terms of a $NH(x,n-1,m)$ distribution so that

$$E\left[\frac{1}{x+y}\right] = \frac{n}{x(m+n)} \quad .$$

Then $E\left[\frac{1}{m+n-x-y}\right]$, for y having a $NH(x,n,m)$ distribution is

$$E\left[\frac{1}{m+n-x-y}\right] = \sum_{y=0}^{m} \frac{\binom{m}{y} \quad (x+y-1)!\,(m+n-x-y-1)!/(m+n)!}{(x-1)!\,(n-x)!/n!}$$

$$= \frac{n}{(m+n)(n-x)} \sum_{y=0}^{m} \frac{\binom{m}{y} \quad (x+y-1)!\,(m+n-1-x)!/(m+n-1)!}{(x-1)!\,(n-1-x)!/(n-1)!}$$

$$= \frac{n}{(m+n)(n-x)}$$

Thus the right member of equation (5.6) is approximately equal to

$$1 + \frac{n}{24N}\left(\frac{1}{x} + \frac{1}{n-x}\right) \quad ,$$

and an approximate normalizing constant for the MLPD is exhibited by the relationship:

$$[1 - \frac{n}{24N}(\frac{1}{x} + \frac{1}{n-x})] \frac{2\pi e^{-N} n!}{N!\,\Gamma(x+\frac{1}{2})\,\Gamma(n-x+\frac{1}{2})} \sum_{y=0}^{m} \binom{m}{y} (x+y)^{x+y} (N-x-y)^{N-x-y}$$

$$\cong 1 \,. \tag{5.7}$$

Using (3.2a) to get an exact expression for the two gamma-functions leads to the alternate relationship

$$[1 - \frac{n}{24N}(\frac{1}{x} + \frac{1}{n-x})] \frac{2^{2n+1} e^{-N} n!\,x!\,(n-x)!}{N!\,(2x)!\,(2n-2x)!} \sum_{y=0}^{m} \binom{m}{y} (x+y)^{x+y} (N-x-y)^{N-x-y}$$

$$\cong 1 \tag{5.8}$$

Finally a simpler expression is obtained by substituting for the gamma functions in (5.7) the approximation (5.5):

$$[1 + \frac{m}{24N}(\frac{1}{x} + \frac{1}{n-x})] \frac{e^{-m} n!}{N!\,x^x (n-x)^{n-x}} \sum_{y=0}^{m} \binom{m}{y} (x+y)^{x+y} (N-x-y)^{N-x-y}$$

$$\cong 1 \tag{5.9}$$

Tables III and IV give the exact value of the left hand side of (5.8) and (5.9) respectively, and this for various values of x, m and n. Note that these developments do not apply for the case $x = 0$.

TABLE III. APPROXIMATION (5.8) - DEVIATION FROM 1.D (*10^2)

N	X	M= 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1	.312	.634	.740	.766	.759	.739	.713	.685	.658	.631	.606	.582	.560	.539	.519
3	1	.087	.306	.406	.452	.471	.476	.472	.464	.454	.442	.430	.418	.406	.395	.383
4	1	-0.006	.160	.282	.341	.374	.392	.401	.404	.403	.399	.394	.388	.381	.374	.367
4	2	.045	.138	.184	.208	.219	.223	.223	.220	.216	.211	.206	.200	.195	.189	.184
5	1	-0.059	.106	.208	.273	.314	.341	.358	.368	.374	.376	.376	.374	.371	.367	.363
5	2	.018	.083	.119	.140	.152	.159	.161	.162	.161	.159	.156	.153	.150	.147	.144
6	1	-0.093	.055	.154	.222	.268	.301	.323	.339	.349	.356	.360	.362	.362	.361	.359
6	2	.004	.056	.089	.109	.122	.130	.135	.137	.138	.138	.137	.136	.134	.132	.130
6	3	.008	.047	.070	.085	.094	.099	.102	.103	.103	.102	.101	.100	.098	.097	.095
7	1	-0.118	.017	.112	.180	.230	.266	.293	.312	.326	.337	.344	.349	.352	.353	.354
7	2	-0.004	.041	.070	.090	.104	.113	.119	.123	.126	.127	.127	.127	.126	.125	.124
7	3	.003	.031	.050	.062	.069	.075	.078	.080	.081	.081	.081	.080	.079	.078	.077
8	1	-0.137	-0.013	.078	.146	.197	.235	.265	.287	.305	.318	.328	.335	.340	.344	.347
8	2	-0.010	.030	.058	.077	.091	.101	.108	.113	.117	.119	.120	.121	.121	.120	.120
8	3	-0.001	.022	.038	.049	.057	.062	.065	.068	.069	.070	.070	.070	.070	.069	.068
8	4	.000	.020	.033	.042	.046	.052	.055	.057	.058	.059	.059	.059	.058	.058	.057
9	1	-0.152	-0.037	.050	.116	.168	.208	.239	.264	.284	.299	.311	.321	.328	.334	.338
9	2	-0.015	.021	.048	.067	.081	.092	.099	.105	.109	.112	.114	.116	.116	.117	.117
9	3	-0.003	.017	.031	.041	.048	.054	.057	.060	.062	.063	.064	.064	.064	.064	.064
9	4	-0.001	.014	.024	.032	.037	.041	.044	.046	.047	.048	.048	.048	.048	.048	.048
10	1	-0.164	-0.057	.026	.091	.142	.183	.216	.242	.264	.281	.295	.306	.315	.323	.328
10	2	-0.018	.015	.040	.059	.073	.084	.092	.098	.103	.107	.109	.111	.113	.114	.114
10	3	-0.005	.013	.026	.035	.042	.048	.052	.055	.057	.058	.059	.060	.060	.060	.060
10	4	-0.002	.010	.019	.026	.031	.035	.037	.039	.040	.041	.042	.042	.043	.043	.042
10	5	-0.002	.009	.018	.023	.028	.031	.033	.035	.036	.037	.038	.038	.038	.038	.038
11	1	-0.174	-0.074	.005	.068	.119	.160	.194	.222	.245	.264	.279	.292	.302	.311	.318
11	2	-0.021	.010	.033	.052	.066	.077	.085	.092	.098	.102	.105	.107	.109	.111	.111
11	3	-0.006	.010	.022	.031	.038	.043	.047	.050	.053	.055	.056	.057	.057	.058	.058
11	4	-0.003	.008	.016	.022	.027	.030	.033	.035	.036	.037	.038	.038	.039	.039	.039
11	5	-0.002	.007	.013	.018	.022	.025	.027	.029	.030	.031	.031	.032	.032	.032	.032
12	1	-0.183	-0.069	-0.013	.048	.098	.140	.174	.203	.227	.247	.263	.277	.269	.299	.307
12	2	-0.023	.005	.028	.045	.060	.071	.080	.087	.093	.097	.101	.104	.106	.108	.109
12	3	-0.007	.007	.019	.027	.034	.040	.044	.047	.050	.052	.053	.054	.055	.056	.056
12	4	-0.004	.006	.013	.019	.023	.027	.029	.031	.033	.034	.035	.036	.036	.036	.037
12	5	-0.003	.005	.011	.015	.019	.021	.023	.025	.026	.027	.028	.028	.028	.029	.029
12	6	-0.002	.005	.010	.014	.017	.020	.022	.023	.024	.025	.025	.026	.026	.026	.026
13	1	-0.190	-0.102	-0.029	.030	.080	.121	.156	.185	.210	.230	.248	.263	.276	.287	.296
13	2	-0.025	.012	.023	.040	.054	.065	.074	.082	.088	.093	.097	.100	.103	.105	.106
13	3	-0.008	.006	.016	.025	.031	.037	.041	.044	.047	.049	.051	.052	.053	.054	.054
13	4	-0.004	.005	.011	.017	.021	.024	.027	.029	.031	.032	.033	.034	.034	.035	.035
13	5	-0.003	.004	.009	.013	.016	.019	.021	.022	.023	.024	.025	.025	.026	.026	.026
13	6	-0.003	.003	.008	.011	.014	.016	.018	.019	.020	.021	.022	.022	.023	.023	.023
14	1	-0.196	-0.113	-0.043	.014	.063	.104	.138	.168	.193	.215	.234	.250	.263	.275	.285
14	2	-0.027	-0.002	.019	.035	.049	.060	.070	.077	.084	.089	.093	.097	.100	.102	.104
14	3	-0.009	.004	.014	.022	.029	.034	.038	.042	.045	.047	.049	.050	.051	.052	.053
14	4	-0.004	.004	.010	.015	.019	.022	.025	.027	.029	.030	.031	.032	.033	.033	.033
14	5	-0.003	.003	.008	.011	.014	.017	.019	.020	.021	.022	.023	.024	.024	.024	.025
14	6	-0.003	.002	.006	.010	.012	.014	.016	.017	.018	.019	.019	.020	.020	.020	.020
14	7	-0.002	.002	.006	.009	.011	.013	.015	.016	.017	.018	.018	.019	.019	.019	.019
15	1	-0.202	-0.123	-0.056	-0.000	.047	.088	.123	.152	.178	.200	.220	.236	.251	.263	.274
15	2	-0.028	-0.005	.015	.031	.045	.056	.065	.073	.079	.085	.089	.093	.096	.099	.101
15	3	-0.009	.002	.012	.020	.026	.032	.036	.039	.042	.045	.047	.048	.050	.051	.052

TABLE IV. APPROXIMATION (5.9) - DEVIATION FROM 1.0 ($*10^2$)

N	X	M= 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1	.826	1.032	1.068	1.048	1.009	.965	.920	.877	.837	.801	.767	.736	.707	.681	.656
3	1	.412	.571	.632	.650	.649	.637	.621	.602	.583	.564	.546	.529	.512	.496	.481
4	1	.293	.437	.510	.547	.563	.567	.565	.558	.550	.539	.529	.517	.506	.495	.484
4	2	.155	.224	.254	.265	.267	.263	.257	.249	.240	.232	.223	.215	.207	.199	.191
5	1	.233	.365	.442	.488	.515	.529	.536	.537	.535	.531	.526	.519	.512	.504	.496
5	2	.096	.146	.172	.185	.190	.191	.189	.185	.181	.176	.171	.166	.161	.156	.151
6	1	.195	.317	.394	.445	.478	.499	.512	.519	.523	.523	.522	.519	.515	.510	.504
6	2	.072	.113	.138	.151	.159	.163	.164	.163	.161	.158	.155	.152	.149	.145	.141
6	3	.054	.084	.101	.110	.115	.116	.116	.115	.113	.110	.108	.105	.102	.099	.096
7	1	.169	.281	.357	.410	.447	.472	.490	.502	.509	.513	.515	.515	.514	.512	.509
7	2	.058	.095	.118	.133	.142	.147	.150	.151	.151	.150	.148	.146	.144	.141	.139
7	3	.037	.060	.074	.082	.087	.089	.090	.090	.069	.088	.087	.085	.093	.081	.079
8	1	.149	.253	.327	.381	.420	.448	.469	.484	.495	.502	.507	.509	.511	.510	.509
8	2	.050	.083	.105	.120	.130	.136	.140	.143	.144	.144	.144	.142	.141	.139	.136
8	3	.029	.048	.060	.068	.073	.076	.078	.078	.078	.078	.077	.076	.074	.073	.072
8	4	.025	.040	.050	.057	.061	.063	.064	.065	.064	.064	.063	.062	.061	.059	.058
9	1	.133	.230	.302	.355	.396	.426	.450	.467	.480	.490	.497	.502	.505	.507	.507
9	2	.044	.074	.095	.110	.121	.128	.133	.137	.139	.140	.140	.140	.139	.138	.137
9	3	.024	.041	.052	.059	.064	.068	.070	.071	.072	.072	.071	.071	.070	.069	.068
9	4	.018	.031	.039	.045	.048	.050	.052	.053	.053	.053	.052	.052	.051	.050	.049
10	1	.121	.211	.280	.333	.374	.406	.431	.451	.466	.478	.486	.493	.498	.501	.504
10	2	.039	.067	.087	.102	.113	.121	.127	.131	.134	.136	.137	.137	.137	.137	.136
10	3	.021	.035	.046	.053	.059	.062	.065	.066	.067	.068	.068	.068	.067	.067	.066
10	4	.015	.025	.032	.037	.041	.043	.045	.046	.046	.046	.046	.046	.045	.045	.044
10	5	.013	.023	.029	.033	.036	.038	.040	.040	.041	.041	.041	.040	.040	.039	.039
11	1	.110	.195	.262	.314	.355	.388	.414	.435	.452	.465	.475	.484	.490	.495	.498
11	2	.035	.061	.081	.096	.107	.115	.122	.126	.130	.132	.134	.135	.135	.135	.135
11	3	.018	.032	.042	.049	.054	.058	.061	.063	.064	.065	.065	.065	.065	.065	.064
11	4	.013	.022	.028	.033	.036	.038	.040	.041	.042	.042	.042	.042	.042	.042	.041
11	5	.010	.018	.023	.027	.030	.032	.033	.034	.034	.034	.035	.034	.034	.034	.033
12	1	.102	.182	.245	.296	.337	.371	.398	.420	.438	.452	.464	.474	.481	.488	.492
12	2	.032	.057	.075	.090	.101	.110	.117	.122	.126	.129	.131	.132	.133	.134	.134
12	3	.017	.029	.038	.045	.050	.054	.057	.060	.061	.062	.063	.063	.064	.064	.063
12	4	.011	.019	.025	.029	.033	.035	.037	.038	.039	.040	.040	.040	.040	.040	.039
12	5	.009	.015	.020	.023	.026	.027	.029	.030	.030	.031	.031	.031	.031	.030	.030
12	6	.008	.014	.018	.021	.024	.025	.026	.027	.028	.028	.028	.028	.028	.028	.028
13	1	.094	.170	.231	.281	.321	.355	.383	.405	.424	.440	.453	.464	.472	.480	.485
13	2	.030	.053	.071	.085	.096	.105	.112	.118	.122	.125	.128	.130	.131	.132	.133
13	3	.015	.027	.035	.042	.048	.052	.055	.057	.059	.060	.061	.062	.062	.062	.062
13	4	.010	.017	.023	.027	.030	.033	.034	.036	.037	.037	.038	.038	.038	.038	.038
13	5	.007	.013	.017	.020	.023	.024	.026	.027	.027	.028	.028	.028	.028	.028	.028
13	6	.006	.011	.015	.018	.020	.021	.022	.023	.024	.024	.024	.024	.024	.024	.024
14	1	.088	.160	.218	.267	.307	.340	.368	.392	.411	.428	.442	.454	.463	.471	.478
14	2	.027	.049	.067	.081	.092	.101	.108	.114	.119	.122	.125	.128	.129	.131	.131
14	3	.014	.025	.033	.040	.045	.049	.052	.055	.057	.058	.060	.060	.061	.061	.061
14	4	.009	.016	.021	.025	.028	.031	.032	.034	.035	.036	.036	.037	.037	.037	.037
14	5	.006	.011	.015	.018	.021	.022	.024	.025	.025	.026	.026	.026	.027	.027	.027
14	6	.005	.010	.013	.015	.017	.019	.020	.020	.021	.021	.022	.022	.022	.022	.022
14	7	.005	.009	.012	.014	.016	.017	.018	.019	.020	.020	.020	.021	.021	.021	.020
15	1	.082	.150	.207	.254	.294	.327	.355	.379	.399	.416	.431	.443	.454	.463	.471
15	2	.026	.046	.063	.077	.088	.097	.104	.110	.115	.119	.122	.125	.127	.129	.130
15	3	.013	.023	.031	.038	.043	.047	.050	.053	.055	.057	.058	.059	.060	.060	.061

V.6. Asymptotic properties

V.6.1 Limits when $m/n \to 0$

The number of observed items becomes infinite, but m is finite. Then it is known that $x/n \overset{a.s.}{\to} p_o$, the true value of the parameter p; consequently if we assume that p_o is neither 0 nor 1, $x \to \infty$ and $n - x \to \infty$ with probability one. The limiting forms of the MLPD and Fisher's likelihood have been given in Section II.3.1. We recall these results:

$$\hat{f}(y;x) \overset{a.s.}{\to} \binom{m}{y} p_o^y q_o^{m-y}$$

and

$$\tilde{L}(y;x) \overset{a.s.}{\to} \frac{p_o^y q_o^{m-y}}{(y/m)^y (1 - y/m)^{m-y}} \quad \text{for} \quad y = 0,1,\ldots,m$$

We now show that the Bayesian predictive distributions for $1/p$ and $1/(1 - p)$ priors have the same limiting form as the MLPD, i.e., tend towards the true distribution of Y, almost surely.

For a $1/p$ prior we have

$$f_1(y|x) \propto \binom{m}{y} (x + y - 1)!\ (n - x + m - y)!\ .$$

Since (with probability 1) x and $n - x$ become infinite we can replace the factorials by Stirling's expression and obtain

$$\lim f_1(y|x) \propto \binom{m}{y} e^{-(x+y-1)} (x+y-1)^{x+y-\frac{1}{2}} e^{-(N-x-y)}$$

$$* \quad (n-x+m-y)^{n-x+m-y+\frac{1}{2}}$$

$$\propto \binom{m}{y} \quad x\left(1+\frac{y-1}{x}\right)^{x+y-\frac{1}{2}}\left[(n-x)\left(1+\frac{m-y}{n-x}\right)\right]^{n-x+m-n+\frac{1}{2}}$$

$$\propto \binom{m}{y} \quad x^y e^{y-1}(n-x)^{m-y} e^{m-y}$$

$$\propto \binom{m}{y} \left(\frac{x}{n}\right)^y \left(1-\frac{x}{n}\right)^{m-y}$$

Finally

$$f_1(y\,|x) \overset{a.s.}{\to} \binom{m}{y} p_o^y q_o^{m-y} \qquad \text{for} \quad y = 0,1,\ldots,m \quad .$$

The proof for a $1/(1-p)$ prior is similar.

V.6.2 Limits when m and $n \to \infty$, m/n remaining constant

We first look at the limiting distribution for Y , where Y has a Negative Hypergeometric distribution with parameters (x,n,m) . The probability $\Pr\{Y = y\}$ can be seen as the probability of obtaining y white balls up until x black balls are drawn from a finite population containing n black balls and m white balls. The NH distribution is more frequently defined for $Z = Y + x$, the total number of balls to be drawn before obtaining x black balls.

Matuzewski (1961) gives the expectation and variance of this distribution and shows that when m and n tend to infinity, m/n remaining constant, it tends to a $NB[x,n/(n+m)]$ distribution. However in the present case $x \to \infty$ and Y tends to have a Normal distribution (see L. N. Bol'shev (1964).) for which we now derive the expectation and variance. For any x,n and m we have

$$E(Y) = \frac{mx}{n+1}$$

$$V(Y) = \frac{mx(n-x+1)(m+n+1)}{(n+2)(n+1)^2}$$

For m and n large these reduce to

$$E(Y) \cong mx/n$$

$$V(Y) \cong m(m+n)x(n-x)/n^3$$

$$\cong m^2(x/n)(1-x/n)(1/m+1/n)$$

That is, Y has a limiting distribution such that

$$\frac{Y/m - \bar{x}}{\sqrt{\bar{x}(1-\bar{x})(\frac{1}{m}+\frac{1}{n})}} \sim N(0,1)$$

Obviously this is also true for Y having a $NH(x+1,n,m)$ distribution and consequently also for Y being distributed according to the MLPD by theorem 5.5 .

We show now that the same applies for the distribution derived from Fisher's Likelihood, with density $\tilde{f}(y;x,n,m)$.

From Lemma 5.4 the function of y ,

$$\frac{y!(m-y-1)!}{y^y(m-y)^{m-y}} ,$$

is an increasing sequence when $y = o,1,\ldots,m-1$, so that

$$\frac{\tilde{f}(y;x,n,m)}{\hat{f}(y;x,n+1,\ m-1)}$$

is also increasing for $y = 0,1,\ldots,m-1$. By defining $\hat{f}(y;x,n+1,m-1)$ equal to zero for $y = m$, one can apply theorem 3.4, i.e.,

$$\tilde{F}(y;x,n,m) = \sum_{k=0}^{y} \tilde{f}(k;x,n,m) \leq \sum_{k=0}^{y} \hat{f}(y;x,n+1,m-1) = \hat{F}(y;x,n+1,m-1) \quad (5.10)$$

Consider now

$$\frac{\hat{f}(y-1;\ x-1,n,\ m-1)}{\tilde{f}(y;x,n,m)} \propto \frac{y^y(m-y)^{m-y}}{(y-1)!(m-y)!} \ .$$

From lemma 5.4 this ratio is increasing for $y = 1,2,\ldots,m,$ so that chosing $\hat{f}(y-1;x-1,n,m-1)$ equal to zero when $y = 0$, we have

$$\tilde{F}(y;x,n,m) = \sum_{k=0}^{y} \tilde{f}(k;x,n,m) \geq \sum_{k=0}^{y-1} \hat{f}(k;x-1,n,m-1)$$

$$\tilde{F}(y;x,n,m) \geq \hat{F}(y;x-1,n,m-1) - \hat{f}(y;x-1,n,m-1)$$

$$\tilde{F}(y;x,n,m) \geq \hat{F}(y;x-1,n,m-1) - \varepsilon \quad (5.11)$$

where $\varepsilon \to 0$ when $x \to \infty$

Finally from (5.10) and (5.11),

$$\hat{F}(y;x-1,n,m-1) - \varepsilon \leq \tilde{F}(y;x,n,m) \leq \hat{F}(y;x,n+1,m-1) \quad ,$$

and the limit is obviously the same for these three CDF's when x,n,m tend to infinity. We formalize the former results in the following theorem.

Theorem 5.6: For $n,m \to \infty$ but m/n constant the predictive CDF's

$F_1(y|x), F_2(y|x), \hat{F}(y;x)$ and $\hat{F}(y;x)$ tend to be such that

$$\frac{Y/m - \bar{x}}{\sqrt{\bar{x}(1-\bar{x})(\frac{1}{m}+\frac{1}{n})}} \xrightarrow{d} N(0,1)$$

with probability one.

V.6.3 Limits when $m \to \infty$, n finite

i) limit for $\hat{f}(y;x)$

When $m \to \infty$ we also have $y \to \infty$ and $m - y \to \infty$ assuming that p_o is not equal to zero or one . Then

$$\hat{f}(y;x) \propto \frac{(x+y)^{x+y}(m+n-x-y)^{m+n-x-y}}{(m-y)!\ y!}$$

and

$$\lim \hat{f}(y;x) \propto y^{x-\frac{1}{2}}(1+\frac{x}{y})^{x+y}(m-y)^{n-x-\frac{1}{2}}(1+\frac{n-x}{m-y})^{m-y+n-x}$$

$$\propto y^{x-\frac{1}{2}}e^{x}(m-y)^{n-x-\frac{1}{2}}e^{n-x}$$

$$\propto y^{x-\frac{1}{2}}(m-y)^{n-x-\frac{1}{2}}$$

Thus the limit in terms of $\bar{y} = y/m$ is ,

$$\hat{f}(\bar{y};x) \to \text{BETA}(x-\tfrac{1}{2};\ n-x-\tfrac{1}{2}) \quad ,$$

so that for m large we can approximately predict $\bar{Y}$ from a BETA distribution.

ii) Limit for $\tilde{f}(y;x)$

$$\tilde{f}(y;x) \propto \frac{(x+y)^{x+y}(N-x-y)^{N-x-y}}{y^y(m-y)^{m-y}}$$

$$\propto y^x\left(1+\frac{x}{y}\right)^{x+y}(m-y)^{n-x}\left(1+\frac{n-x}{m-y}\right)^{m-y+n-x}$$

$$\propto y^x(m-y)^{n-x}$$

i.e.,

$$\tilde{f}(y;x) \to \text{BETA}(x,n-x)$$

iii) Limit for the Bayesian predictions

Let us consider the limiting form of a $NH(x,n,m)$ when only $m \to \infty$

$$f_Z(z) \propto \binom{m}{y} (m+n-x-y)!\ (x+y-1)!$$

$$\lim f_Z(z) \propto \frac{(m+n-x-z)^{m+n-x-z+\frac{1}{2}}(x+z-1)^{x+z-\frac{1}{2}}}{z^{z+\frac{1}{2}}\ (m-z)^{m-z+\frac{1}{2}}}$$

$$\propto \frac{\left[(m-z)\left(1+\frac{n-x}{m-z}\right)\right]^{m-z+n-x+\frac{1}{2}} z^{z+x-\frac{1}{2}}\left(1+\frac{x-1}{z}\right)^{z+x-\frac{1}{2}}}{z^{z+\frac{1}{2}}\ (m-z)^{m-z+\frac{1}{2}}}$$

$$\propto (m-z)^{n-x}\left(1+\frac{n-x}{m-z}\right)^{m-z+n-x+\frac{1}{2}} z^{x-1}\left(1+\frac{x-1}{z}\right)^{z+x-\frac{1}{2}}$$

$$\propto \quad (m - z)^{n - x} z^{x - 1} \quad .$$

So that the limiting distribution for a NH(x,n,m) when $m \to \infty$ is a BETA(x - 1, n - x) distribution. By replacing x by x + 1 for a NH(x + 1,n,m) we get a BETA(x,n - x - 1) distribution.

We summarize the results in the table below.

Approximations for m large ($\bar{y} = y/m$)

Type of Approach	Notation	Limiting Distribution	Limiting Expectation
Bayesian 1/p	$f_1(\bar{y}\,\vert x)$	BETA(x - 1,n - x)	$\frac{x}{n+1}$
Bayesian 1/1 - p	$f_2(\bar{y}\,\vert x)$	BETA(x,n - x - 1)	$\frac{x+1}{n+1}$
MLPD	$\hat{f}(\bar{y};x)$	BETA(x - ½,n - x - ½)	$\frac{x+\frac{1}{2}}{n+1}$
FISHER	$\hat{f}(\bar{y};x)$	BETA(x,n - x)	$\frac{x+1}{n+2}$

V.7 Exact probabilities of coverage associated with the MLPD.

For any family of prediction intervals $[L(X),U(X)]$ the confidence level β has been defined such that

$$\Pr_{\theta}\{L(X) \leq Y \leq U(X)\} = \beta ,$$

where this probability is independent of the parameter θ. When the prediction intervals are obtained from the MLPD, the probability above will usually depend on the value of the parameter. In this section we propose to show the variation of the exact probability of coverage as a function of the parameter p for the binomial distribution when inferences are made from the MLPD.

Once a nominal level of confidence (say .90) has been chosen and $X = x$ has been observed the MLPD is used as an ordinary distribution to compute a lower limit $L_{.95}(x)$ and an upper limit $U_{.95}(x)$ for a prediction interval. Actually, since we are dealing with a discrete distribution, these limits are randomized in the usual way. We proceed now to the computation of the probability of coverage for a given value p_o of the parameter p. We have

$$\begin{aligned}\beta(p_o) &= \Pr_{p_o}\{L_{.95}(X) \leq Y \leq U_{.95}(X)\} \\ &= \sum_{x=0}^{n} \Pr_{p_o}\{L_{.95}(x) \leq Y \leq U_{.95}(x) \mid X = x\} \Pr_{p_o}\{X = x\} \\ &= \sum_{x=0}^{n} \Pr_{p_o}\{L_{.95}(x) \leq Y \leq U_{.95}(x)\} \Pr_{p_o}\{X = x\}\end{aligned}$$

Figures 5.1,5.2, and 5.3 show how the function $\beta(p)$ deviates from the nominal value for $0 < p < 1$. The computations are only for upper one-sided intervals because curves for the corresponding lower one-sided intervals can be obtained by symmetry around $p = \frac{1}{2}$. For instance

$$Pr_p\{Y \le U_{.95}(X)\} = Pr_{1-p}\{L_{.95}(X) \le Y\}$$

Figures 5.1 and 5.3 exhibit the improvement occurring when the size, n, of the observed sample increases for nominal levels .95 and .99 respectively. Theorem 2.9 guarantees that these curves will tend to the horizontal line at the nominal value when n tends to infinity. In comparing figures 5.1 and 5.3 we also notice that the higher the nominal confidence level is, the smaller are the deviations from that level. In other words the prediction will be more reliable for higher nominal levels.

Figure 5.2 shows the changes induced by the increase of the size of the future sample; for large m the curve has lower minima.

Overall we can see that the procedure tends to be anti-conservative for all values of p except around 0. Because of the symmetry property for lower limits, the underestimation of the true level takes place for p around 1. Consequently for a two-sided interval the curve will be symmetric with respect to $p = \frac{1}{2}$ with a flat relative maximum around $p = \frac{1}{2}$ located below the nominal value.

EXACT CONFIDENCE LEVELS OF MLPD-INTERVALS

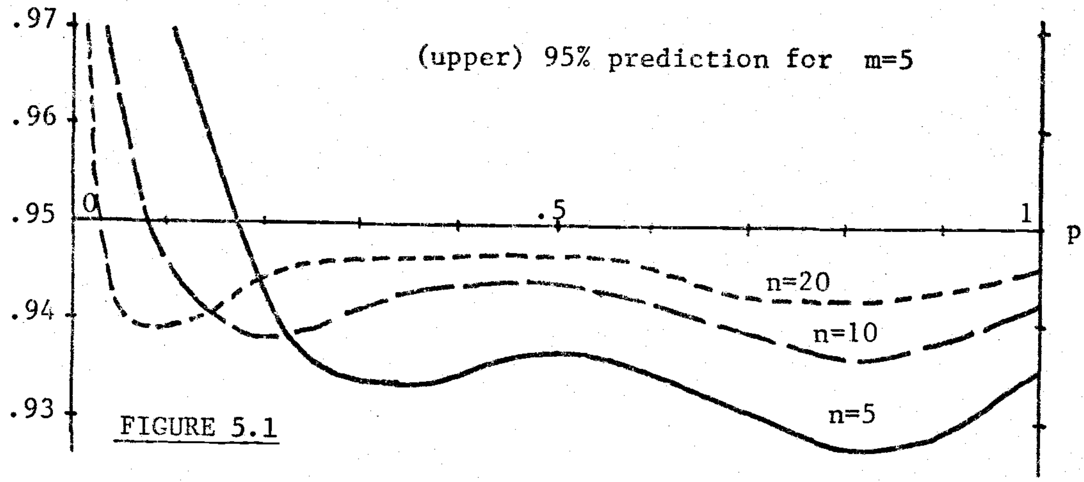


FIGURE 5.1

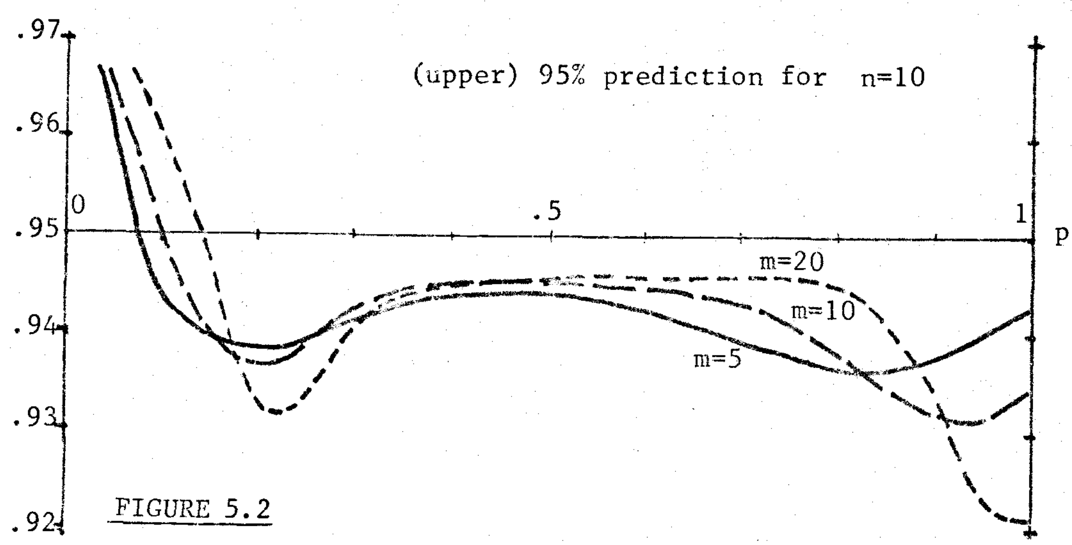


FIGURE 5.2

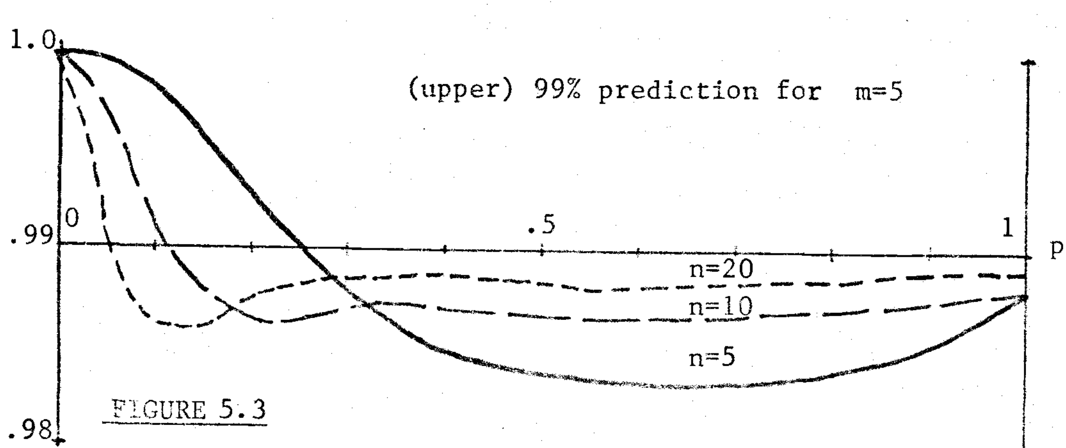


FIGURE 5.3

V.8 Examples

The eight following examples are exhibited:

$$\begin{array}{lll} n = 5 & m = 5 & x = 1,2 \\ & m = 10 & x = 1,2 \\ n = 10 & m = 5 & x = 2,4 \\ & m = 10 & x = 2,4 \end{array} \quad .$$

Because of the symmetry in $x, n - x$ along with $y, m - y$ and $p, 1 - p$ only small values of x are considered.

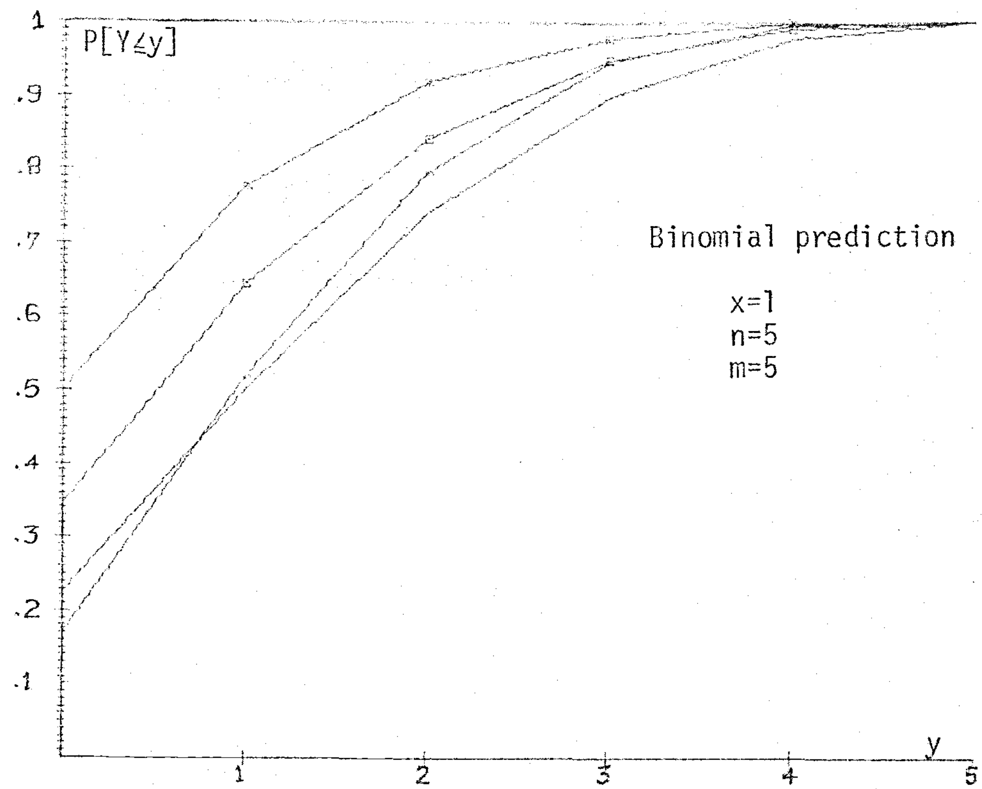
For each example four predictive CDF's are plotted:

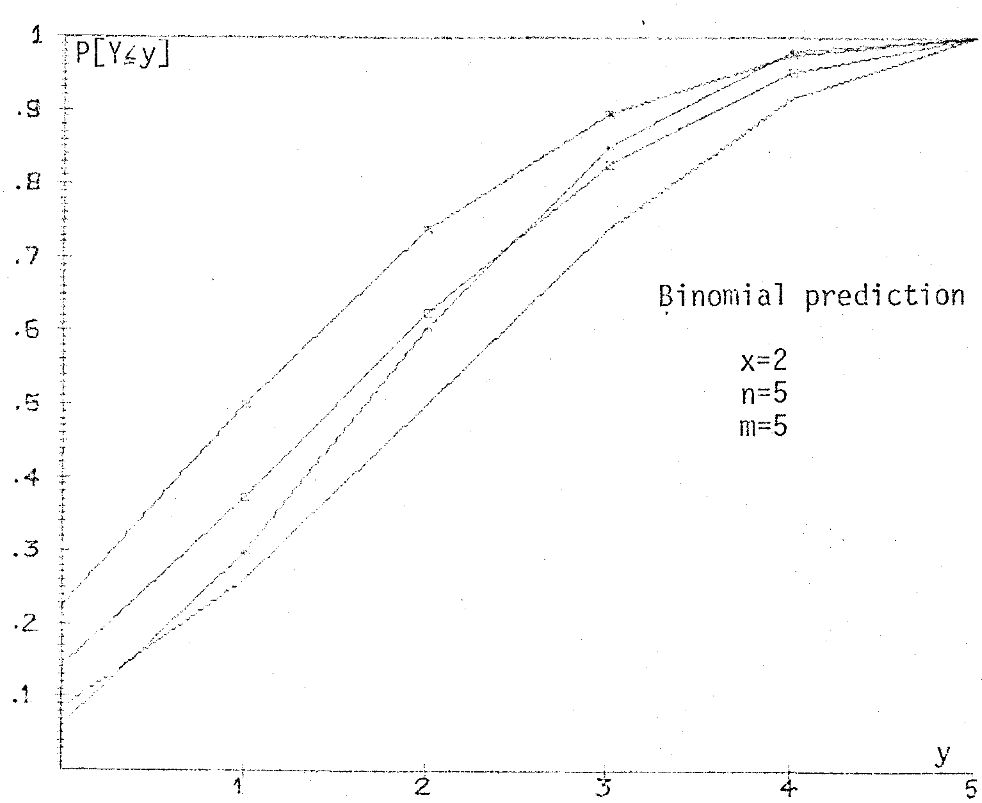
i) Bayesian with $1/p$ prior (marked *)

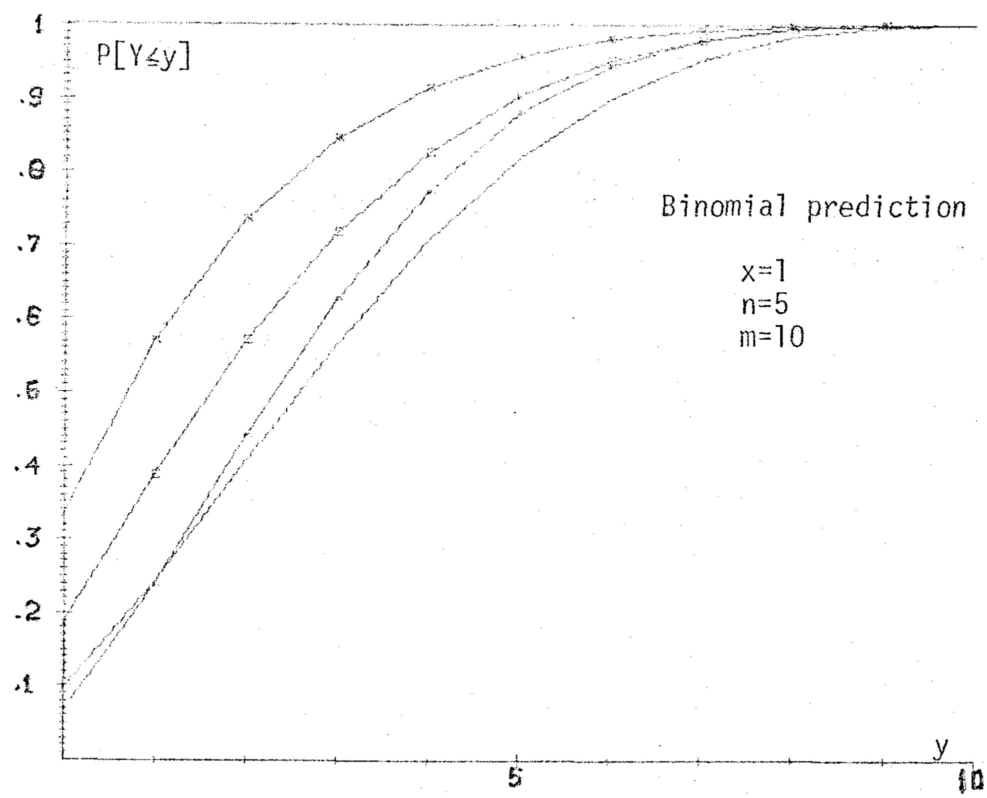
ii) MLPD (marked ◘)

iii) Bayesian with $1/(1 - p)$ prior (no marks)

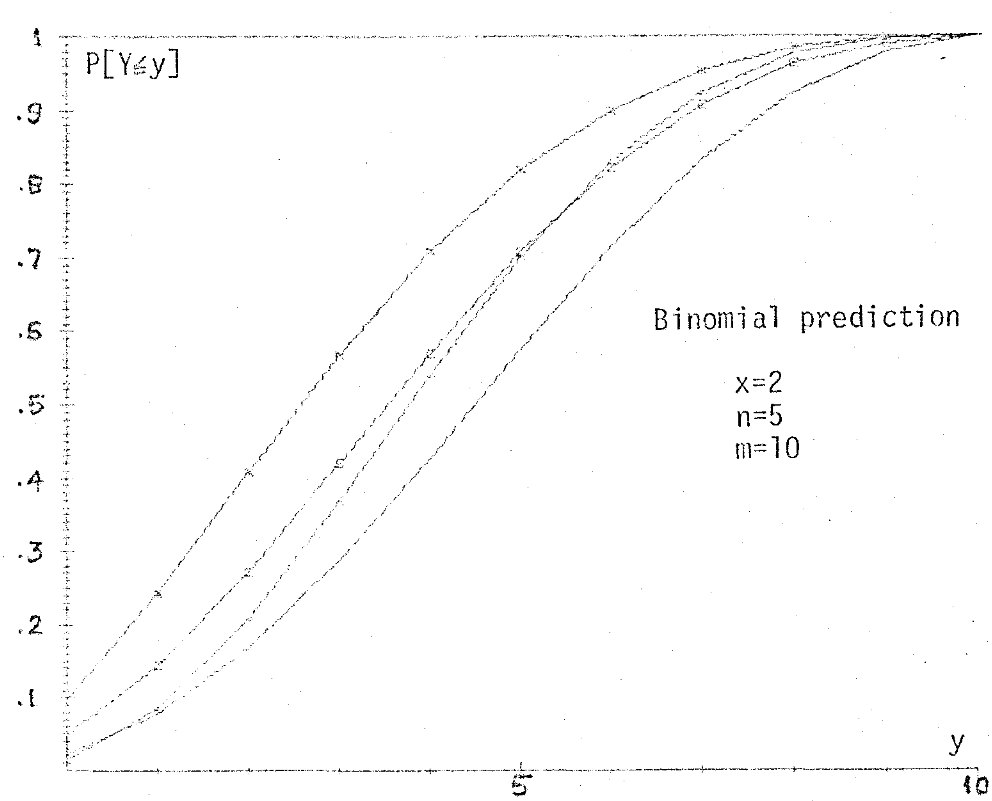
iv) Fisher's approach (marked ▲)


P[Y≤y]
Binomial prediction
x=1
n=5
m=5
y
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
1
2
3
4
5

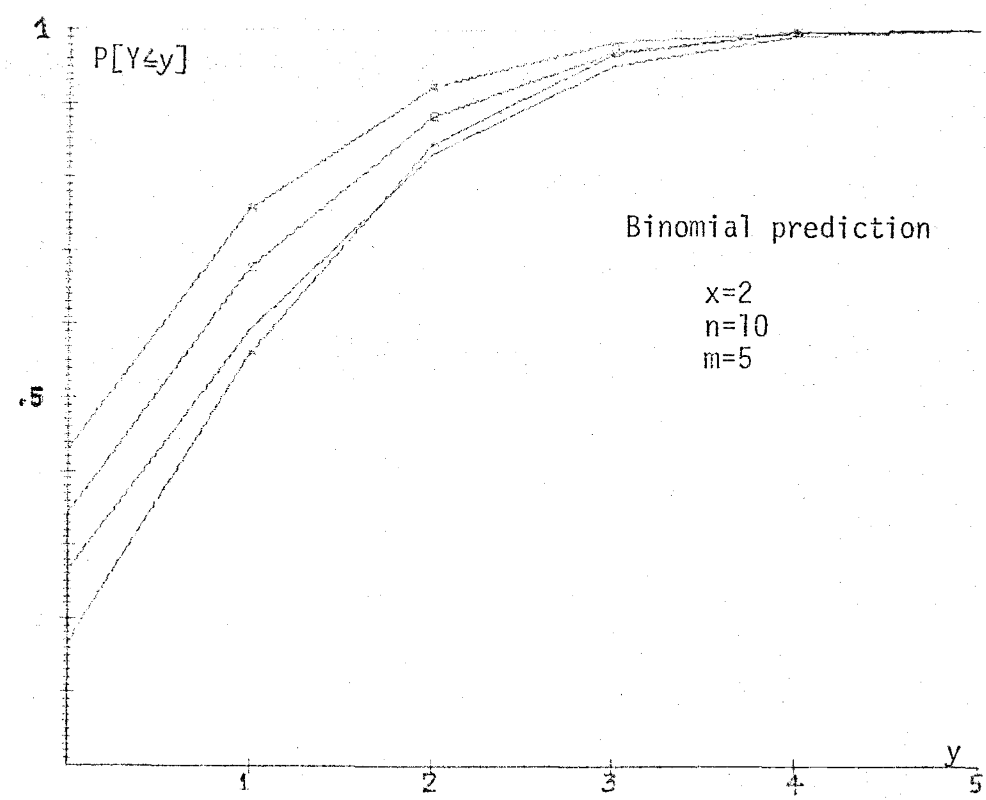

P[Y≤y]
Binomial prediction
x=2
n=5
m=5
y
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
1
2
3
4
5

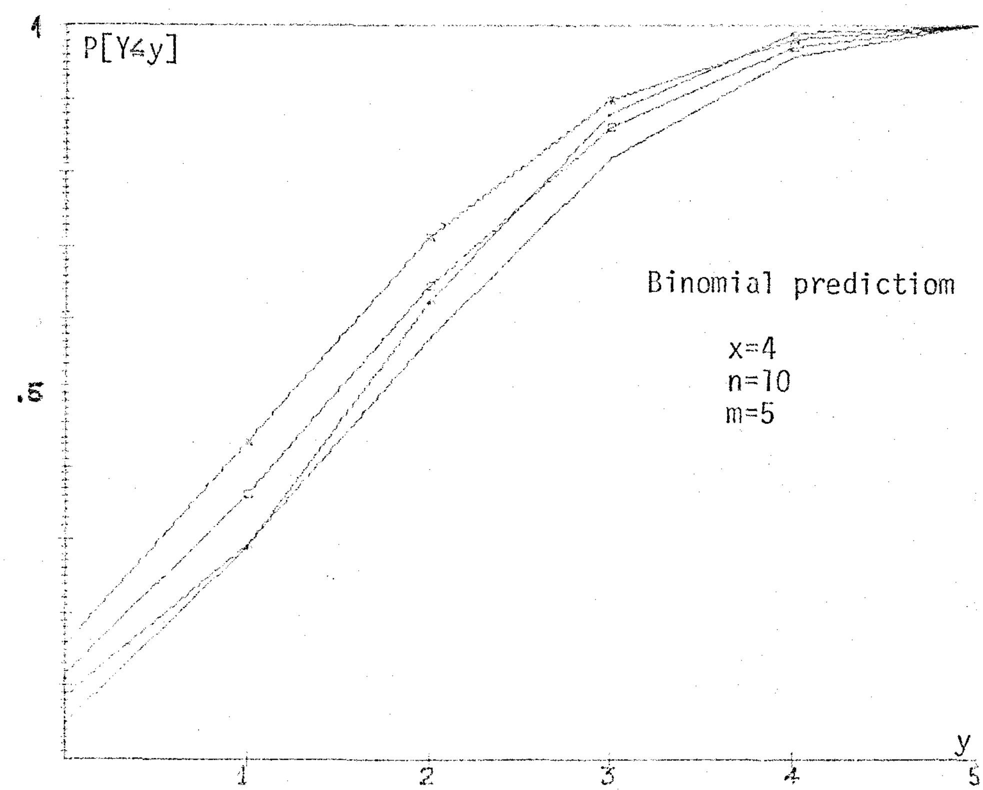


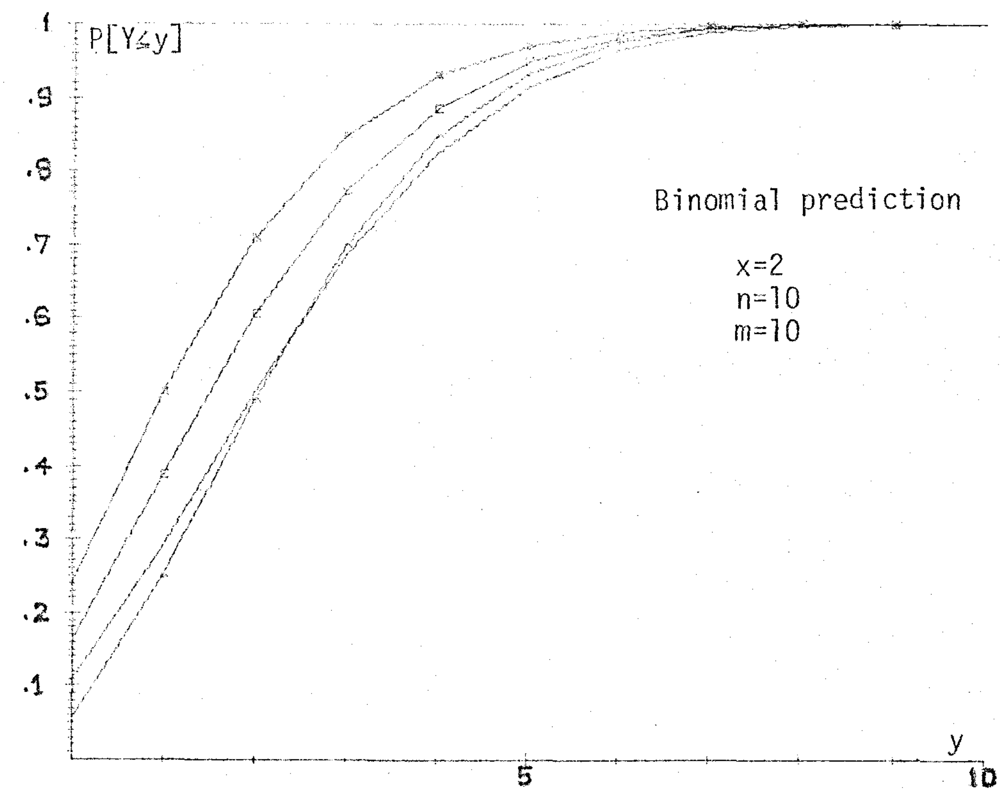
1
P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Binomial prediction
x=1
n=5
m=10
y
5
10

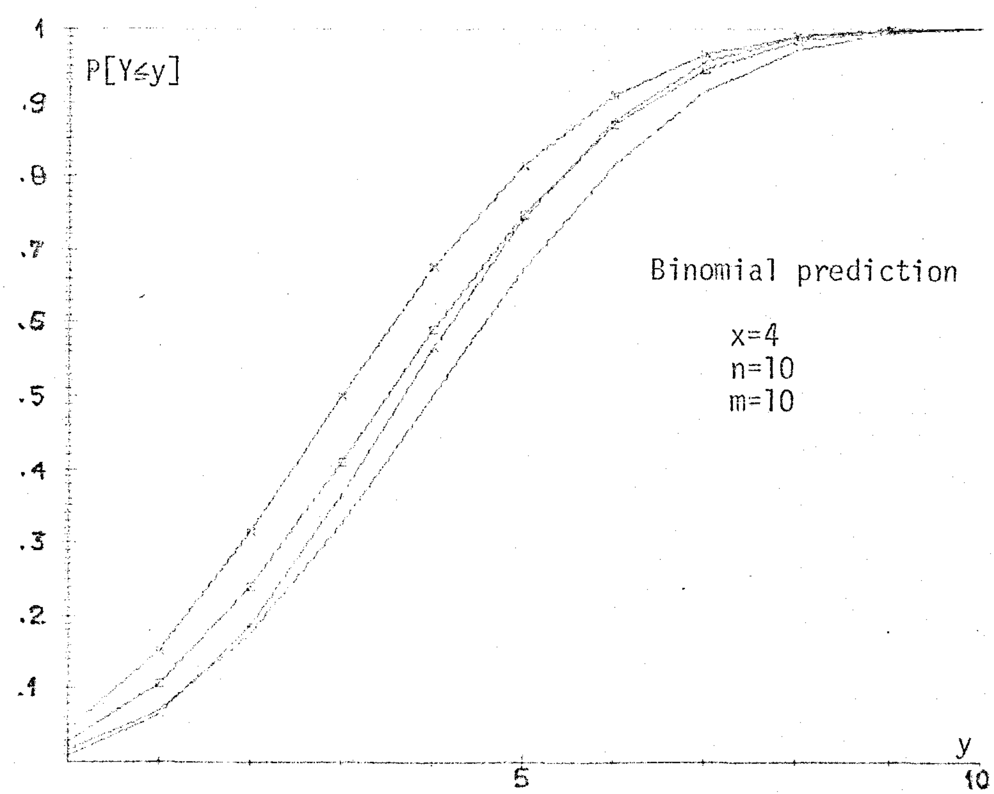


1
P[Y≤y]
.9
.8
.7
.6
.5
.4
.3
.2
.1
Binomial prediction
x=2
n=5
m=10
y
5
10


P[Y≤y]
Binomial prediction
x=2
n=10
m=5
1
.5
y
1
2
3
4
5


P[Y≤y]
Binomial predictiom
x=4
n=10
m=5
1
.5
y
1
2
3
4
5


P[Y≤y]
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
Binomial prediction
x=2
n=10
m=10
y
5
10


P[Y≤y]
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
Binomial prediction
x=4
n=10
m=10
y
5
10

V.9 Conclusions.

As in Poisson sampling the MLPD lies between a conservative and an anticonservative distribution. Thus the true probability of coverage for a prediction interval is deviating little from the nominal value of the confidence level. For example, the deviation is no larger than one percent for a one-sided 99% interval.

The three distributions of interest are $NH(x + \frac{1}{2},n,m)$ (approximately), $NH(x,n,m)$ and $NH(x + 1,n,m)$. These correspond to the familiar $p(1 - p)^{-\frac{1}{2}}$, p^{-1} and $(1 - p)^{-1}$ Bayes priors respectively.

It is no surprise that the results for the Poisson and the Binomial sampling have so many similarities, since the first is a limiting form of the second. In fact the prediction problems for these two cases can be related in the following way.

Suppose $X \sim B(n,p)$ and $Y \sim B(m,p)$, and let n and m tend to infinity with n/m remaining constant. Let also p tend to zero in such a manner that $np \to \lambda$, then:

$$X \xrightarrow{d} P(\lambda)$$

$$Y \xrightarrow{d} P(\lambda m/n) \quad ,$$

i.e., X and Y can be seen as the sample sums of two samples from the same Poisson population with sample sizes having the ratio n/m (recall that the Poisson prediction problem depends on the sample sizes only through their ratio).

Then a $NH(x,n,m)$ distribution becomes a $NB(x,n/(m + n))$ distribution, which is a result given by Matuzewski (1962). The Bayesian priors $1/p$, or equivalently $1/(np)$, and $1/(1 - p)$ become $1/\lambda$ and uniform priors respectively.

VI STRATIFIED RANDOM SAMPLING FROM BINOMIAL POPULATIONS.

We turn now to the same problem as in chapter IV except that we are dealing here with binomial populations instead of Poisson populations. Because the variable to be predicted, i.e., the grand sum of k samples, has no explicit density function no prediction can be made without carrying out numerical computations of the convolution type.

VI.1 Problem and solutions.

We assume that we are sampling n_i items from stratum i and we observe x_i successes. Let p_i be the unknown probability of success for stratum i.

The binomial problem is more complex than the Poisson problem because the sum of binomial random variables with distinct p_i's does not have an explicit mass probability function.

Therefore in the frequentist framework we are not able to exhibit a function $\psi(y,x_1,x_2,\ldots,x_k)$ such that the distribution of Y given ψ would be independent of the p_i's .

In the Bayesian approach when the priors are of the form $p_i^{\alpha_i}(1 - p_i)^{\beta_i}$ we have to sum negative hypergeometric random variables. There is also no explicit form of the distribution of such a sum but there is no obstacle to derive it in a computational manner.

By analogy with the Poisson case one might feel that choosing a prior $1/p_i$ for each stratum (resp. $1/(1 - p_i)$) might be too extreme, so that one would think of using more moderate priors such as

$$\frac{1}{p_i(1 - p_i)^{(k-1)/k}} \quad \text{or} \quad \frac{1}{p_i^{(k-1)/k}(1 - p_i)} \tag{6.1}$$

that correspond to the $1/\lambda_i$ and $(1/\lambda_i)^{(k-1)/k}$ priors for the Poisson problem.

Unfortunately there is no special case like in Poisson sampling to justify from a frequentist point of view the choice of certain priors. However some justification may be found in the choice of (6.1) by showing that, when the strata tend to be identical, these priors yield predictive distributions that tend to agree respectively with the $1/p$ and $1/(1 - p)$ priors of simple random sampling.

If the strata were identical then inferences would be made from a $NH(x, \sum m_i, \sum n_i)$ and a $NH(x + 1, \sum m_i, \sum n_i)$ distribution respectively, where $x = \sum_{i=1}^{k} x_i$. The expectation of the grand total Y would be

$$\frac{\sum m_i}{\sum n_i + 1} \sum x_i \qquad \text{and} \qquad \frac{\sum m_i}{\sum n_i + 1} (\sum x_i + 1) \tag{6.2}$$

Now suppose we pick for each stratum i a prior of the type $p_i^{\alpha}(1 - p_i)^{\beta}$. Then the posterior distribution for Y_i would be $NH(x_i + \alpha + 1, n_i + \alpha + \beta + 1, m_i)$ (see section V.3) whose expectation is

$$\frac{m_i(x_i + \alpha + 1)}{n_i + \alpha + \beta + 2}$$

Thus the expectation of Y would be

$$\sum_{i=1}^{k} \frac{m_i}{n_i + \alpha + \beta + 2} (x_i + \alpha + 1) \quad . \tag{6.3}$$

In order to compare this value to the values in (6.2) we need to restrict the special case further to $m_i = m$ and $n_i = n$ for all i. Then (6.2) becomes

$$\frac{m}{n + 1/k} (\sum x_i) \quad \text{and} \quad \frac{m}{n + 1/k} (\sum x_i + 1) \quad .$$

The only way in (6.3) to chose α and β in order to have the same expectations for both approaches is in fact as exhibited in (6.1).

Note that when the number of strata k increases the two classes of priors of (6.1) tend to be confounded and equal to $[p_i(1 - p_i)]^{-1}$, for which

$$E(Y) = \sum_{i=1} m_i x_i / n_i \quad .$$

Because of the absence of an explicit density function for Y, the prediction likelihood approaches also fail to apply. However one can think of obtaining a predictive distribution by convoluting the predictive distributions of each stratum, i.e.,

$$\hat{f}(y; x_1, x_2, \ldots, x_k) \cong \sum_{\Sigma y_i = y} \hat{f}(y_i; x_i) \quad .$$

The expectation for this distribution is

$$E(Y) \cong \sum_{i=1}^{k} m_i(x_i + \tfrac{1}{2})/n_i$$

It can be seen in the following examples how this result compares with some Bayesian predictive distributions.

VI.2 Examples

In the next pages two predictive densities and CDF's are given for the following approaches:

i) Bayesian under priors $1/p_1$, $1/p_2$ (marked *)

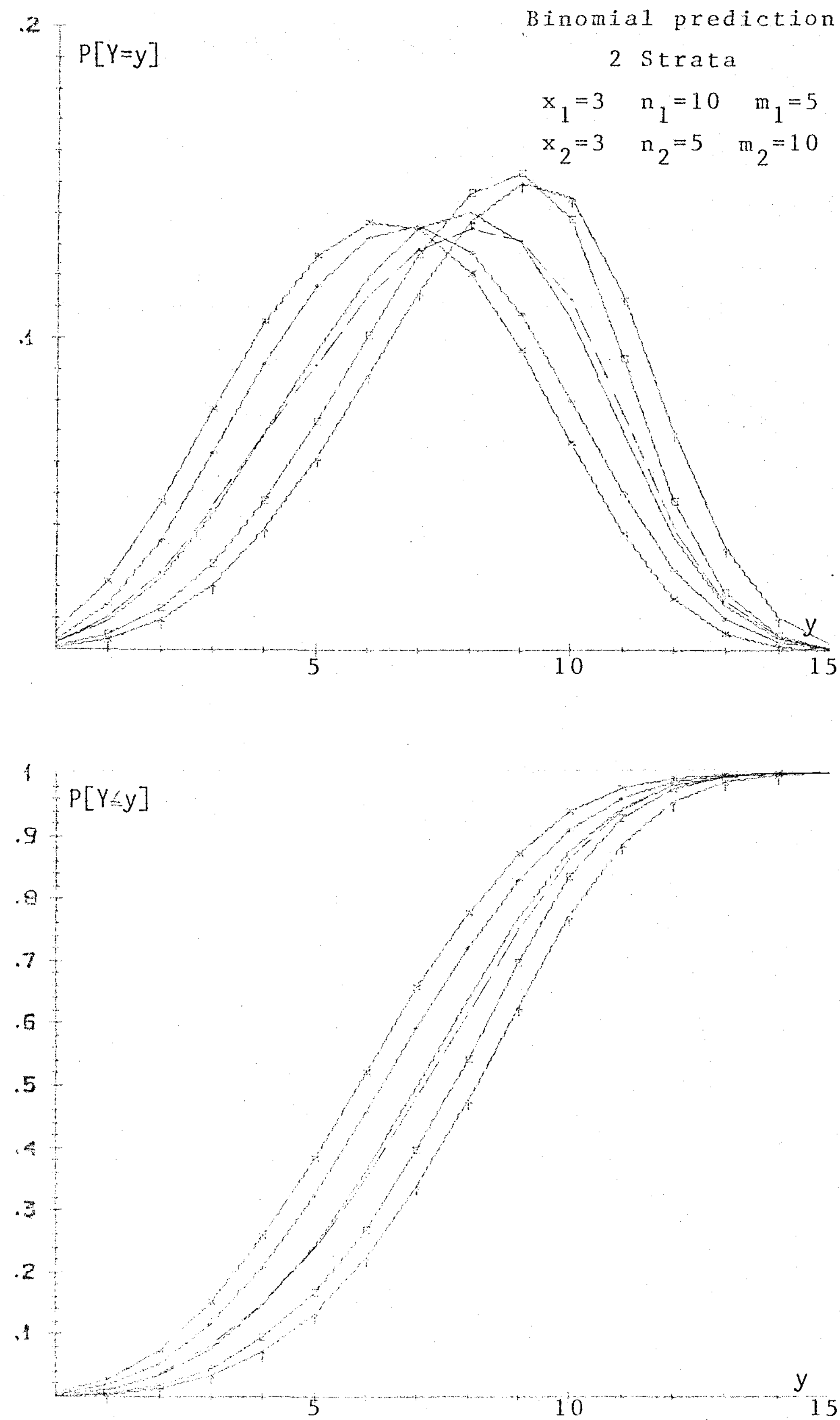
ii) Bayesian under $1/(1 - p_1)$, $1/p_2$ (marked ▲)

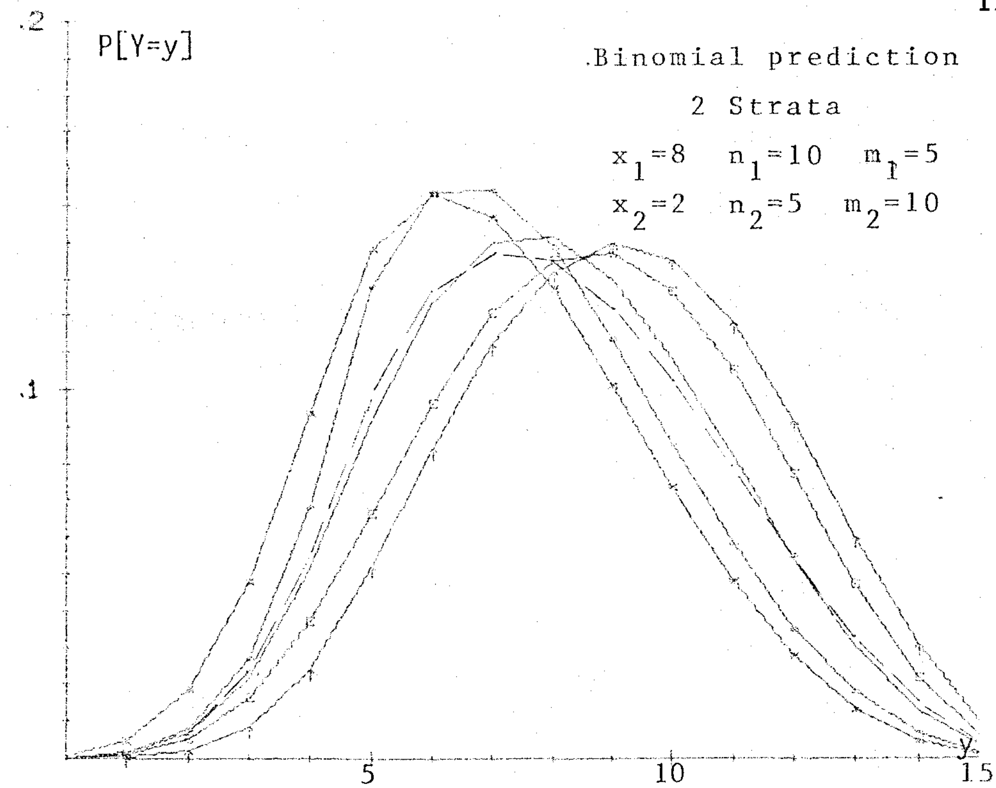
iii) Convolution $(MLPD)_1 * (MLPD)_2$ or approximately Bayesian $[p_1(1 - p_1)]^{-\frac{1}{2}}$, $[p_2(1 - p_2)]^{-\frac{1}{2}}$ (no marks)

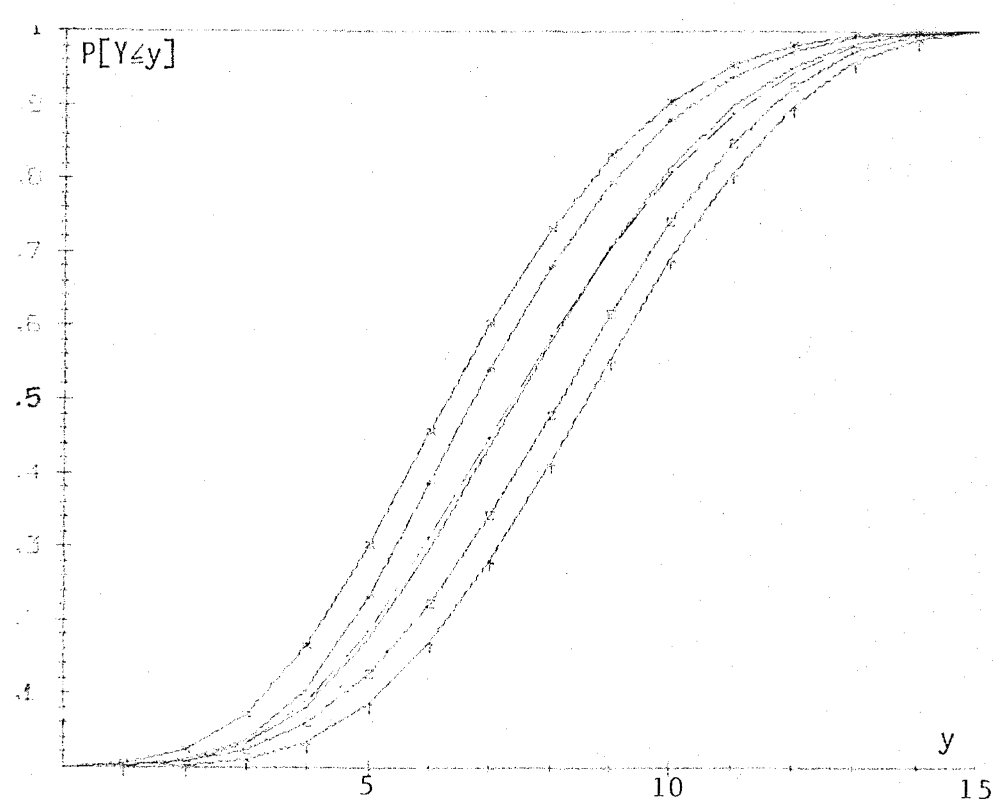
iv) Bayesian under $1/p_1$, $1/(1 - p_2)$ (marked ◘)

v) Bayesian under $1/(1 - p_1)$, $1/(1 - p_2)$ (marked †)

vi) Bayesian under $[p_1(1 - p_1)]^{-1}$, $[p_2(1 - p_2)]^{-1}$ (dashes)


Binomial prediction
2 Strata
$x_1=3$ $n_1=10$ $m_1=5$
$x_2=3$ $n_2=5$ $m_2=10$
P[Y=y]
.2
.1
5
10
15
y
P[Y≤y]
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
y


.2
P[Y=y]
Binomial prediction
2 Strata
x1=8 n1=10 m1=5
x2=2 n2=5 m2=10
.1
5
10
15
y


P[Y≤y]
.5
5
10
15
y

VII GENERAL CONCLUSIONS AND COMMENTS

It has been shown in this thesis that the Maximum Likelihood Predictive Distribution behaves quite satisfactorily for binomial and Poisson predictions (and to a lesser extent for the Normal prediction).

For the Bayesian statistician who would compare prospective values of y on the basis of some "weighted integration" of the likelihood function, $L(\theta;y|\underset{\sim}{x})$, over θ for each y, the question will be how well the maximum value of this likelihood function can stand for the whole function. For instance when the likelihoods remain identical for all y values except for a proportionality constant on the abcissa (which occurs in the normal prediction of a sample mean when the variance is known), then comparing the likelihoods through their maxima is equivalent to comparing them through their area, that is, the MLPD is the Bayesian distribution under uniform prior. In any case one would like the likelihood function to shift enbloc along the θ axis when y varies, which analytically is somewhat expressed by the monotone likelihood ratio condition for the density family of y.

In the binomial and Poisson cases these requirements are fulfilled and from the Bayesian point of view it is reassuring that the MLPD's are close approximations to the Bayesian predictive distributions obtained by using priors $[p(1 - p)]^{-\frac{1}{2}}$ and $\lambda^{-\frac{1}{2}}$ respectively. In fact the MLPD would be the first approximation of these Bayesian predictions if an expression were derived that would be easy to manage analytically or computationally.

We have been able to show the strong consistency of the MLPD, which incidentally is uniform in y for common cases. One can feel intuitively that Bayesian predictive distributions will have the same properties, but as far as I know this has not been proven yet, and I suspect a general proof to be a rather sophisticated task.

Note that the consistency by itself is a weak requirement although a necessary one. In fact if we simply choose as a predictive distribution the member of the parametric family of distributions for y which corresponds to a value of the parameter equal to the maximum likelihood estimator based on the observations, we see that by continuity this predictor is also consistent. However, one does not expect it to perform well, at least as far as prediction limits on y are concerned, since it entails only one source of uncertainty. Therefore other criteria should be taken into account such as expected loss under some standard loss functions, or expected bias.

In this thesis there has not been any consideration of optimality for the choice of a family of prediction intervals among others. The choice between Bayesian and non-Bayesian is a conceptual one whereas the choice among non-Bayesian procedures would be based on criteria such as minimal interval length or uniform most accuracy.

The intervals inferred from the MLPD are not truly prediction intervals in the sense that there is no a-priori guarantee that the probability of coverage will be at least as large as the given level. Nevertheless we saw for the binomial case in section V.7 that starting with a nominal level of 99%, for instance, we obtained a procedure guaranteeing a true level of confidence of at least 98% (recall figure V.3). Also some prior knowledge about the unknown parameter will make results of the type of those in section V.7 more useful. At any rate further developments of this type of study would shed more light on the comparison of different methods; in particular the results of section V.7 should be extended to the Bayesian predictions under priors $1/p$ and $1/(1-p)$ in order to see how conservative or anticonservative these really are. The deficiencies of Normal approximations could also be evaluated in this way.

Finally the author feels that the MLPD is a useful outsider in prediction theory just as the maximum likelihood estimator in estimation theory, performing quite satisfactorily in some instances, poorly in others and turning up as the most tractable solution to some complex cases.

BIBLIOGRAPHY

Aitchison,J. and Sculthorpe, D. (1965). Some problems of statistical prediction. Biometrika 52,469-483.

Barnard,G.A. (1969). Summary remarks. New Developments in Survey Sampling,699-707. New-York. Wiley-Interscience.

Bol'shev,L.N. (1964). Distributions related to the hypergeometric distribution. Theory of Prob. and its Appl. 9,619-623.

Borel,E. (1942). Sur l'emploi du theoreme de Bernoulli pour faciliter le calcul d'une infinite de coefficients.Application au probleme de l'attente a un guichet. C. R. Acad. Sci. Paris,214,452-456.

Doetsch,G. (1950). Handbuch der Laplace-Transformation.Vol.1- Basel. Verlag Birkhaueser.

Faulkenberry,G.D. (1972). Some results on prediction intervals. Technical Report Nr.14,Dpt. of Statistics,Oregon State University.

Faulkenberry,G.D. (1973). A method of obtaining prediction intervals. J. Amer. Stat. Assoc. 68,433-435.

Feller,W. (1957). An Introduction to Probability Theory and Its Applications,Vol.1,Sec.Ed. New-York. John Wiley and Sons Inc.

Fisher,R.A. (1959). Statistical Methods and Scientific Inferences,Sec. Ed. London. Oliver and Boyd.

Haight,F.A. and Breuer,M.A. (1960). The Borel-Tanner distribution. Biometrika 47,143-146.

Hudson,D.J. (1971). Interval estimation from the likelihood function. J. Royal.Stat. Soc. 33,256-262.

Johnson,N.L. and Kotz,S. (1969). Discrete Distributions. Boston. Houghton Mifflin Co.

Kalbfleisch,J.D. (1971). Likelihood methods of prediction. Foundations of Statistical Inference, A Symposium,378-392. Toronto,Montreal. Holt,Rinehart and Wilson of Canada Ltd.

Kalbfleish,J.D. and Sprott,D.A. (1969). Applications of likelihood and fiducial probability to sampling finite populations. New Developments in Survey Sampling,358-389. N.Y. Wiley-Interscience.

Kempthorne,O. (1969). Some remarks on statistical inference in finite sampling. New Developments in Survey Sampling,680-686. New-York Wiley-Interscience.

Matuzewski,T.I. (1962). Some properties of Pascal distribution for finite population. J. Amer. Stat. Assoc. 57,172-174.

Nelson,W. (1970). Confidence intervals for the ratio of two Poisson means and Poisson prediction intervals. IEEE Transactions on Reliability (May 1970),42-49.

Olsen,D.E. (1974). Statistical Prediction Intervals. PhD Thesis, Oregon State University.

Pratt,J.W. (1965). Bayesian interpretation of standard inference statements. J. Royal Stat. Soc. B,27,169-203.

Stevens,W.L. (1950). Fiducial limits for the parameter of a discontinuous distribution. Biometrika 37,117-129.

Thatcher,A.R. (1964). Relationships between Bayesian and confidence limits for predictions. J. Royal Stat. Soc. B 26,176-192.

Widder,D.V. (1961). Advanced Calculus,Sec.Ed. Englewoods Cliff N.J. . Prentice Hall Inc.

Widder,D.V. (1971). An Introduction to Transformation Theory. New-York Academic Press.

Zacks,S. (1971). The Theory of Statistical Inference. New-York. John Wiley and Sons Inc.