Thailand Statistician
January 2021; 19(1): 42-57
http://statassoc.or.th
Contributed paper

Asymptotic Analysis of Method of Moments Estimators of Both Parameters for the Binomial Distribution: Theoretical Part

Salma Saad*[a], Shakhawat Hossain [b] and Andrei Volodin [a]
[a] Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada.
[b] Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Canada.
*Corresponding author; e-mail: saad202s@uregina.ca

Received: 27 July 2019
Revised: 14 September 2019
Accepted: 10 October 2019

Abstract

Estimating the both unknown parameters m (number of trails) and p (success probability) of the binomial distribution on the basis of fixed sample size n has been unsolved over the years. Many questions regarding asymptotic distribution for small or large sample properties of the estimatos have been ignored or have received inadequate treatment. The aim of this paper is to study the asymptotic properties of estimators for the binomial distribution by the method of moments. The asymptotic normality of the joint estimators is estiblished using the delta method.

Keywords: Delta method, asymptotic normality, hypothesis testing, confidence sets.

1. Introduction

An estimation of parameters of the binomial distribution using a sample of fixed size n, when both parameters m (number of triels) and p (success probability) are unknown, has remained an important statistical problem for more than three quarters of a century. This can be explained by the difficulty of finding a solution for this problem and the fact that a relatively small number of easily computed rigorous estimators exist that are robust in the presence of outliers. Known estimates of m usually underestimate the true value of m. An estimate of the parameter m has many interesting and important practical applications, such as a count of the number of mistakes (bugs) in computer codes (see, for example, DasGupta and Rubin 2005), and detection of the size of a closed population of animals. One of the most important applications that we are planning to conser in futher papers is the detection of the number of synaptic vesicles with acetylcholine in a nerve ending for investigating a neuromuscular junction (synapse) (Nicholls et al. 2001). Therefore, a derivation of estimators for the parameters m and p of the binomial distribution is an important problem to solve.

The main difficulty of this problem with these estimators that derived directly by the method of moments is that they do not have moments of all orders. Also, it is possible to obtain negative values of the estimator for the parameter p and values of the estimator for m that are smaller than the largest value in the sample. We are mostly interested in studing the properties of these estimators as no solution of this problem was presented in the literature.

There are the four general methods for the construction of estimators of parameters of binomial distribution: method of maximum likelihood, method of moments, method of least squares, and Bayesian method. In this paper, we consider only the method of moments for the estimation of parameters m and p of the binomial distribution. Since these estimators do not have moments of all orders, we can not obtained the mean, variance, and covariance for these estimators. Therefore, we apply the delta method to derive the asymptotic normality of the joint distribution of the method of moments estimators of m and p.

The asymptotic normality of the estimators cannot be interpreted directly as characteristics of accuracy properties of these estimators, since the estimators do not have mean and varaince. Therefore, we suggest a modification of the estimators based on adding a constant term to the denominators of the estimators. It is better to say that a constant is added in the formulae of statistics that are defined by the estimators. We prove that this modification has no influence on the parameters of the established asymptotic normality. For the modified estimates, the asymptotic normality becomes the characteristics of mean value, variance and covariance for the estimators of m and p.

The main tool we are using to derive our theoretical results is the delta method. This method is mentioned in nearly all textbooks in mathematical statistics, for example, the classical mathematical statistics book by Cramér (2016) or a more modern textbooks by van der Vaart (1998) and Lehmann (2004).

Throughout this paper we consider the Binomial distribution with parameters m (number of trails) and p (success probability), where m is a positive integer and $0 \le p \le 1$. The Binomial distribution is denoted $Bin(m, p)$, , and, if a random variable X has a binomial distribution, then its probability mass function is defined as

$$P(X = k) = \binom{m}{k} p^k (1-p)^{m-k}, \quad k = 0, 1, \ldots, m.$$

The paper is organized as follows. In Section 2, we derive formulae for the estimators of the parameter of binomial distribution using the method of moments and derive the joint asymptotic normality in Theorem 3.1 of Section 3. Modified and corrected estimators are introduced in Section 3, and their joint asymptotic normality is shown in Theorem 3.2. The interval estimation and hypothesis testing techniques are discussed in Section 4.

Now we present some literature review. One of the first theoretically rigorous publications devoted to the statistical problems of the binomial distribution was published by Haldane (1941). This paper mentioned a number of practical situations that arise on different occasions when an event occurs a given number of times. Obviously, the amounts of times that the event occurs can be considered as discrete random variables with values in the set of positive integers. A Poisson distribution sometimes gives a good fit for these events; however, in some cases, a good fit may be obtained by the binomial distribution $Bin(m, p)$. Therefore, it is important to study the parameter estimation of the binomial distribution.

According to Haldane (1941), the binomial distribution with the two parameters m and p can always be estimated, which is described as fitted, by the first two moments using the method of moments. If one denotes $\bar{X}$ and S^2 as the sample mean and sample variance, respectively, then it is easy to obtain the Method of Moments estimators of m and p (see Proposition 2.1 in Section 3 for their derivations). Unfortunately, these estimators cannot be considered absolutely efficient (see the discussion right after Proposition 2.1 in Section 3 of this paper). Hence, Haldane (1941) suggested

the use the maximum likelihood estimators. Unfortunately, the maximum likelihood estimators have even more irregular behaviour.

The maximum likelihood estimators for the Binomial distribution have been considered prior to Haldane (1941). Published in 1939, nearly eighty years ago, Jeffreys' famous "Theory of Probability'' textbook pointed out that the maximum likelihood estimators (fitting) seem to require the use of tables of the digamma function, see below. We were unable to find a copy of the first, 1939 edition of the Jeffreys' textbook, but it was reprinted in Jeffreys (1998). The argument about the use of the digamma function tables seems to be minor in our modern computer era, but in the first part of the previous century their use was a serious disadvantage. Because of that, Haldane (1941) argued that the use of digamma function tables may be avoided, and only elementary operations are needed to estimate both parameters of the binomial distribution by the method of maximum likelihood. In our paper we don't consider the maximum likelihood estimators. From our point of view, the main disadvantage of these estimators is that they cannot be derived in closed form (as described in Section 1.1 below).

A more detailed investigation of the maximum likelihood estimators can be found in the paper of Binet (1952). He studied asymptotic properties of these estimators and the method of moments estimator and also mentioned that the estimation results are subject to a variety of errors because both methods give unstable estimators when the sample mean transcended the sample variance. In addition, Binet (1952) considered Hammersley (1950)'s idea of estimating restricted parameters while estimating the number of trials and likelihood function is under specific restrictions or constraints. Skellam (1948) also supported this idea of restrictions mentioned that the estimation can be made more realistic by fitting a distribution with a limited range. Another useful discussion about the maximum likelihood method was given by DasGupta and Rubin (2005).

More information about the accuracy properties of both methods of moments and maximum likelihood is given in Olkin et al. (1981). The paper is devoted to the estimation of the number of trials m of the binomial experiment, given a sequence of independent success counts obtained by replicating the m-trial experiment. This is a less studied and considerably harder problem than estimating the probability of success p for the binomial distribution. Both estimation methods for m are considered in the paper Olkin et al. (1981). Difficulty arises when the mean and the variance estimated by the method of moments are almost equal. Therefore, they considered that both methods exist if and only if the sample mean exceeds the sample variance, specifically when $\frac{\bar{X}}{S^2} \geq 1+\frac{1}{\sqrt{2}} \approx 1.71$. Another unstable case of these estimators occurs when p is small and m is large. For example, consider $m = 75,\ p = 0.32,$ and the observed values (success counts) 16, 18, 22, 25 and 27. Even though p is not small, the method of moments estimate of it from the observed counts is 0.21, quite different from the true value 0.32. This is an example of an unstable case with $\bar{X} \approx S^2,$ even though $\frac{E\bar{X}}{ES^2} = 1.84.$

In Olkin et al. (1981), the idea of stabilizing the method of moments estimators is to use the ridge tracing method proposed by Hoerl and Kennard (1970). This method suggests adding a constant $\varepsilon > 0$ to each observed success count $X_i,$ so that $\bar{X} = \bar{X} + \varepsilon = x,$ while S^2 stays unchanged. Then, the method of moments estimate of the parameter m takes the form $m(x) = x^2/(x - S^2),$ and $m(x)$ reaches a minimum at $x = 2S^2$ and then increases to infinity. Eventually, even though $m(x)$ is

rapidly changing for x near S^2, it stabilizes in some sense at a value denoted as ε. To achieve a stable version of the method of moments estimate, Olkin et al. (1981) suggest an algorithm based on the value of ε. They choose $\varepsilon > 0$ so that $x > S^2$. For consistency, it requires $m(x) \geq X_{\max} + \varepsilon$, where $X_{\max} = \max\{X_1, \ldots X_m\}$; this is because $m(x)$ represents the number of trials, whereas $X_{\max} + \varepsilon$ represents the maximum number of successes in the data set. Their performance described in terms of results obtained by simulation.

In a pioneering paper by Hoel (1947), a test statistic for testing a simple hypothesis $m = m_0$ against a composite alternative $m > m_0$ suggested when the value of the parameter p is unknown. This statistic applied for the construction of a confidence interval for m and a point estimator of m similar to the estimator by the method of moments suggested.

The asymptotic and robustness properties of all kinds of estimators of the parameters of the Binomial distribution are presented in a paper by Hall (1994). The statement by Olkin et al. (1981), Carroll and Lombard (1985), and subsequent authors that the maximum likelihood and method of moments can behave in extremely unstable ways is supported by findings in Hall (1994). Both methods are problematic because they are highly sensitive to small changes in the observed data, which can cause a drastic change in the estimation of the parameters. For example, consider a sample consisting of the observations 16, 18, 22, 25, and 27, where the maximum likelihood and method of moments estimators of m are, respectively, 99 and 102. If we misreport values or small changes happen (say 27 is recorded as 28) in this sample, the corresponding estimators by using the same methods for the new sample are 190 and 195, respectively. Hence we can conclude that, incidentally, the maximum likelihood and method of moments estimators fail to be robust in the presence of registry errors. This case was noted by Olkin et al. (1981).The reasons for this irregular behavior are not clear in the literature. For instance, allowing both numbers of the binomial observations n (sample size) and the true value of m (number of trails) to increase together results in an asymptotically normal distribution for both the maximum likelihood and method of moment estimators of m.

In Hall (1994)'s paper, an alternative asymptotic theory for estimators of m has been developed by allowing the value of m and the number of binomial data value n to increase together while permitting the value of p to decrease at the same time. This is not the case in our research, as we consider m and p fixed and staying the same for all sample sizes n.

DasGupta and Rubin (2005) provided a good survey of estimation methods for the binomial distribution in 2005. They mentioned that estimation of the parameters of the binomial distribution is still an important and attractive problem for at least four reasons:

(1) it is known to be a principally difficult problem, with underestimation of m a serious practical impediment;

(2) even with easily computable and easily motivated estimates, the parameters are still generally lacking;

(3) the problem displays an ingrained instability with the common estimates of m being vulnerable to enormous fluctuations under slight perturbations of one or two sample values; and

(4) it has many interesting practical applications, from species variety estimation to error counting in software codes. In addition, estimating p when m is unknown is also an interesting problem and has actually received less attention in the literature than the corresponding problem to estimate the parameter m.

We completely agree with these points and also consider the problem of binomial distribution parameters estimation as an important one.

Fisher (1941) argued that if we consider sample $X_1, X_2, \ldots, X_n$ from the binomial distribution, then the sample maximum $X_{\max} \to m$ almost surely. Hence, the extreme statistic $X_{\max} = m$ almost surely for all large n. Because of that, Fisher (1941) concluded that the problem of estimation of the number of trials m for the Binomial distribution is not a very interesting one. Although this is a technically correct statement, DasGupta and Rubin (2005) showed that this process is almost incomplete for this technical fact by giving a formula of the smallest sample size n for which $P\left(X_{\max} \geq \frac{m}{2}\right) \geq 1-\alpha,$ where $0 < \alpha < 1.$ For example, the smallest n such that $P\left(X_{\max} \geq \frac{m}{2}\right) \geq 0.5$ is 31,500 when the true m is 100 and the true p is 0.3. Thus, DasGupta and Rubin (2005) showed that the maximum of the observations $X_{\max}$ is not a reliable estimate of m. They also proved that $X_{\max}$ is a biased estimate of $m,$ and quantified how badly biased the maximum of sample observations is an estimator of m.

Some papers such as one by Feldman and Fox (1968) estimated only parameter m and considered p is known. Their papers divided into two parts: the first part dealt with the estimation of the number of trials m by using a maximum likelihood estimate under some specific conditions; the second part considers that m is large so that the $Bin(m, p)$ distribution can be approximated well by normal distribution $N\left(mp, mp(1-p)\right)$ with mean mp and variance $mp(1-p)$ by defining another variable $Y_i = \frac{X_i}{1-p}$. Here, of course, $X_1, X_2, \ldots, X_n$ is a sample from $Bin(m, p)$ distribution. Then Y_i's are approximately normally $N(\mu, \mu)$ distributed, where $\mu = mp/(1-p).$ Three estimators of μ are suggested: the minimum variance unbiased estimate, the maximum likelihood estimator and $\bar{Y}$. However, studying these estimators does not seem to be particularly relevant to their papers' topics or to the problem of the interest, which is to estimate the parameter m of the Binomial distribution.

Other papers, such as those by Draper and Guttman (1971), Raftery (1988), and Carroll and Lombard (1985), consider the Bayesian estimators of the binomial parameters.

The Bayesian approach was adopted by Draper and Guttman (1971) for independent priors distributions for m and $p.$ The prior distribution of parameter m is the discrete uniform (it is assumed that the upper bound for m is known) and the prior for p is the beta. Blumenthal and Dahiya (1981) derived m^* as an estimator of $m,$ where (m^*, p^*) is the joint posterior mode of (m, p) with the above mentioned Draper-Guttman priors; see also Raftery (1988). However, they did not say how the parameters of the beta prior for p should be chosen. Carroll and Lombard (1985) recommended as an m estimator the posterior mode of m with the Draper-Guttman prior after integrating out 0. They called this $Mbeta(2,2)$ estimator.

Raftery (1988) adopted a hierarchical Bayesian approach to estimate the parameters of the Binomial distribution. His method provides a simple and flexible way to specify the prior information. The investigation is mostly focused on the point estimator of $m,$ while interval estimation has been less considered. It is assumed that the prior distribution of m is Poisson and the

case of vague prior is also considered. We do not consider the Bayesian approach in our paper, although it is an interesting topic for future research.

Before we proceed with the material presented in the next sections, we need to discuss some classical and well known definitions and procedures of statistical inference. The main purpose of this is to introduce the notations we are using throughout paper.

1.1. Estimating both parameters of binomial distribution by the method of maximum likelihood

First, we note that the method of moments, not the method of maximum likelihood, is investigated in this paper. However, the method of maximum likelihood is one of the most popular estimation methods, so we briefly present what has been done in this direction.

Let $\{X_i, 1 \le i \le n\}$ be independent identically distributed random variables from the binomial distribution with parameters m and p, and n be the sample size. The probability mass function for the binomial distribution is

$$P(X = k) = \binom{m}{k} p^k (1-p)^{m-k}, \quad k = 0, 1, \ldots, m,$$

and the likelihood function is

$$L(m, p \mid X_i = k_i) = \prod_{i=1}^{n} \binom{m}{k_i} p^{k_i} (1-p)^{m-k_i},$$

where $0 \le k_i \le m$ and $1 \le i \le n$.

Recall the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ for $\alpha > 0$. The derivative of the logarithm of the gamma function is called the digamma function, which can be writen as

$$\psi(\alpha) = (\log \Gamma(\alpha))' = \frac{\Gamma(\alpha)}{\Gamma'(\alpha)},$$

where log is the natural logarithm.

Using the well known property that $\Gamma(k) = (k-1)!$ for any natural k, we can rewrite the binomial coefficients

$$\binom{m}{k} = \frac{m!}{(m-k)!k!} = \frac{\Gamma(m+1)}{\Gamma(m-k+1)\Gamma(k+1)}.$$

Hence, the log-likelihood function for binomial distribution is

$$\begin{aligned} l(m, p \mid X_i = k_i) &= \sum_{i=1}^{n} \log L(m, p \mid X_i = k_i) \\ &= \sum_{i=1}^{n} \left[\log \binom{m}{k_i} + k_i \log p + (m - k_i) \log(1-p) \right] \\ &= \sum_{i=1}^{n} \left[\log \Gamma(m+1) - \log \Gamma(m-k_i+1) - \log \Gamma(k_i+1) + k_i \log p + (m-k_i) \log(1-p) \right]. \end{aligned}$$

In order to find the maximum likelihood estimators, we need to take partial derivatives with respect to the parameters and equate them to zero. There are no difficulties with the parameter p,

$$\frac{\partial l(m, p \mid X_i = k_i)}{\partial p} = \frac{\sum_{i=1}^{n} k_i}{p} - \frac{mn - \sum_{i=1}^{n} k_i}{1-p} = 0.$$

Then the maximum likelihood estimator of p is $\hat{p} = \sum_{i=1}^{n} k_i / mn$. For a derivative by parameter m the situation is far more complicated because it is a discrete parameter that takes only positive integer values. By considering m as a continuous parameter, we are going to take the differentiation of log-likelihood with respect to m and equating the result to zero;

$$\frac{\partial l(m, p \mid X_i = k_i)}{\partial m} = \sum_{i=1}^{n} \left[\psi(m+1) - \psi(m - k_i + 1) + \log(1-p) \right] = 0.$$

As we can see, contrary to the method of moments estimators, there is no closed form for the maximum likelihood estimator of the number of trails. We also see that the maximum likelihood equation involves digamma function, the disadvantage mentioned in Jeffreys (1998).

Haldane (1941) suggested the following way around the difficulty with digamma function. Note that

$$\binom{m}{k} = \frac{m!}{(m-k)!k!} = \frac{\prod_{j=1}^{k-1}(m-j)}{k!},$$

and hence the log-likelihood function takes the form

$$l(m, p \mid X_i = k_i) = \sum_{i=1}^{n} \left[\sum_{j=1}^{k-1} \log(m-j) - \log(k_i!) + k_i \log p + (m - k_i) \log(1-p) \right].$$

Again, by considering as a continuous parameter, we are going to take the differentiation with respect to and equating the result to zero,

$$\frac{\partial l(m, p \mid X_i = k_i)}{\partial m} = \sum_{i=1}^{n} \left[\sum_{j=1}^{k-1} \frac{1}{m-j} + \log(1-p) \right] = 0.$$

There is no way to write a solution of this equation in closed form. DeRiggi (1983) showed that a maximum likelihood estimate exists for both parameters of the binomial distribution if and only if the sample mean exceeds the sample variance. In this case the log-likelihood function can be written as $l\left(m, \sum_{i=1}^{n} k_i / m\right)$ and the first step for analysis is to note that $l\left(m, \sum_{i=1}^{n} k_i / m\right)$ is greater than $l(m, p)$ for any $0 \le p \le 1$. This standard result can be derived by taking the derivative with respect to p, see Binet (1952).

2. Estimation by the Method of Moments and Some Disadvantages of These Estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the binomial $Bin(m, p)$ distribution. It is well known that the sample mean and sample variance

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2$$

are unbiased estimators of the true mean value mp and true variance $mp(1-p)$ of this distribution.

Proposition 2.1 *The estimators of the parameters m and p by the method of moments are*

$$\hat{p}_n = \frac{\bar{X} - S^2}{\bar{X}} \quad \text{and} \quad \hat{m}_n = \frac{\bar{X}^2}{\bar{X} - S^2}.$$

Proof: The estimators of the parameters m and p by the method of moments are the solutions of the system of equations

$$mp = \bar{X}, \quad mp(1-p) = S^2.$$

Simple arithmetic concludes the proof.

Unfortunately, the range of values for these estimators is beyond the natural parametric space of the binomial distribution. The estimator $\hat{p}_n$ may happen to be negative, hence it is impossible to interpret it as a success probability for a binomial experiment. The estimator $\hat{m}_n$ is usually not an integer and it may happen that it is less than the largest number in the sample $X_{\max}$, and may even happen to be negative.

To show the disadvantages of these estimators we can consider an example of a sample of size $n = 2$ from the Binomial distribution with parameters m and $p = 1/2$. Assume the sample values $x_1 = k$ and $x_2 = 0$, where the values of k and m, $0 \le k \le m$ will be specified later. Note that

$$P(X_1 = x_1 = k, X_2 = x_2 = 0) = \binom{m}{k}\left(\frac{1}{2}\right)^{2m} > 0.$$

Note also that $\bar{x} = k/2$ and $s^2 = k^2/4$. Therefore, $\hat{p}_n = 1 - \frac{k}{2}$ and $\hat{m}_n = \frac{k}{2-k}$.

1. The estimator may happen to be negative; hence, it is impossible to interpret it as a success probability for a Binomial experiment. For example, take and in the example above we get with positive probability $\binom{4}{3}(1/2)^{2\times 4} = \frac{1}{64}$.

2. It may happen that the estimator $\hat{m}_n$ is negative. For example, $m = 6$ and $k = 5$ in the example above to get $\hat{m}_n = -\frac{5}{3}$ with positive probability $\binom{6}{5}\left(\frac{1}{2}\right)^{2\times 6} = \frac{3}{2048}$.

3. The estimator $\hat{m}_n$ is usually not an integer.

4. It may happen that the estimator $\hat{m}_n$ is less than the largest number $X_{\max}$ in the sample.

The estimators $\hat{p}_n$ and $\hat{m}_n$ have one more extremely unpleasant property: their mean values and variances do not exist. This is related to the fact that the denominators of the statistics can take zero values with positive probabilities.

In order to explain the absence of mean and variance, consider the following simple argument. Let $X_1, \ldots, X_n$ be a sample from $Bin(p, m)$ population. That is, $X_1, \ldots, X_n$ are independent identically distributed random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ having the $Bin(p, m)$ distribution. Consider the event that all observations are zeros, that is,

$$A = \{\omega \in \Omega : X_1(\omega) = 0,\ X_2(\omega) = 0, \ldots, X_n(\omega) = 0\}.$$

Note that the probability of this event is $P(A) = (1-p)^n > 0$. Also note that if $\omega \in A$, then the sample mean $\bar{X}(\omega) = 0$ and sample variance $S^2(\omega) = 0$, that is $\bar{X}(\omega) = S^2(\omega)$.

It may also happen that $\overline{X}(\omega) = S^2(\omega)$ for other values of sample, not only when all of them are zeros. We did not try to solve the equation $\overline{X}(\omega) = S^2(\omega)$ and find all sample values that satisfy this equation. It is enough for us to say the following. Let event $B = \left\{\omega \in \Omega : \overline{X}(\omega) = S^2(\omega)\right\}$. Then $A \subset B$ and hence $P(B) \geq P(A) > 0$.

This happens especially often when p is small. To show this, we generate samples from binomial distribution with fixed number of trials $m = 10,$ but we vary the probability of the success $p = 0.001 - 0.6,$ and we record the number of cases when $\overline{X} = S^2$ for each p.

3. Asymptotic Normality of the Estimators

The theorem on asymptotic normality of a function of a sample moments does not require any additional assumptions except its differentiability. This allows us to establish the joint asymptotic normality of the estimates $\hat{p}_n$ and $\hat{m}_n,$ not looking at the fact that the mean values and variances of the method of moments estimators $\hat{p}_n$ and $\hat{m}_n$ do not exist. We also note that the direct interpretation of the delta method as an instrument for obtaining asymptotic values for the mean value and variance of method of moments estimators is not appropriate, but it has perfect sense for the modified estimators.

3.1. Method of moment estimators

Lemma 3.1 *For the sampling from the* $Bin(m, p)$ *distribution, the centered sample mean* $\overline{X} - mp$ *and sample variance* $S^2 - mp(1-p)$ *are asymptotically normal* $N(\vec{0}, \tilde{\Sigma})$ *with zero mean vector* $\vec{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ *and covariance matrix*

$$\tilde{\Sigma} = \begin{pmatrix} (mp(1-p))^2 & mp(1-p)(1-2p) \\ mp(1-p)(1-2p) & 2m^2p^2(1-p)^2 + mp(1-p)\big(1-6p(1-p)\big) \end{pmatrix}.$$

Proof: As it is shown in Saad (2019b) Theorem 1 (or Saad 2019a, Theorem 2.1) that the sample mean and sample variance are asymptotically normal, if the fourth moment exists. Namely, if we denote mean $\mu = EX_1,$ variance $\sigma^2 = E(X_1 - \mu)^2,$ the third central moment $\mu_3 = E(X_1 - \mu)^3,$ and the fourth central moment $\mu_4 = E(X_1 - \mu)^4$ then $\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix}$ converges in distribution as $n \to \infty$ to the bivariate normal distribution $N(\vec{0}, \Sigma),$ where $\vec{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$.

It is simple to derive that four central moments of the binomial distribution $Bin(m, p)$ (see Saad 2019a, Lemma 3.1) are as follows

$$E\left(X\right) = mp,\ E(X - EX)^2 = mp(1-p),\ E(X - EX)^3 = mp(1-p)(1-2p),$$

$$E(X - EX)^4 = 3m^2p^2(1-p)^2 + mp\left(1-p\right)\left(1-6p(1-p)\right).$$

Therefore, the sample mean and sample variance are asymptotically normal with the mean and covariance matrix mentioned in the formulation of the lemma.

Theorem 3.1 *The distribution of* $\sqrt{n}\begin{pmatrix}\hat{p}_n - p\\ \hat{m}_n - m\end{pmatrix}$ *for* $n \to \infty$ *converges to the bivariate normal distribution with the zero mean vector and covariance matrix*

$$\Sigma = \begin{pmatrix}\sigma_p^2 & \rho\sigma_p\sigma_m\\ \rho\sigma_p\sigma_m & \sigma_m^2\end{pmatrix},$$

where

$$\sigma_p^2 = \frac{(1-p)(a+p)}{m}, \quad \sigma_m^2 = \frac{m(1-p)a}{p^2}, \rho = -\sqrt{\frac{a}{a+p}},$$

and $a = 2(1-p)(m-1).$

Proof: The delta method consists of expansion of the functions $\hat{p}_n$ and $\hat{m}_n$ into two dimensional Taylor series expansion by the powers of $t_1 = \bar{X}$ and $t_2 = S^2$ in the neighbourhood of mathematical expectations of these statistics, and keeping only linear terms.

Note that in our case

$$\hat{p}_n = \frac{\bar{X} - S^2}{\bar{X}} \equiv g_1(\bar{X}, S^2), \quad \text{that is,} \quad g_1(t_1, t_2) = \frac{t_1 - t_2}{t_1},$$

and

$$\hat{m}_n = \frac{\bar{X}^2}{\bar{X} - S^2} \equiv g_2(\bar{X}, S^2), \quad \text{that is,} \quad g_2(t_1, t_2) = \frac{t_1^2}{t_1 - t_2}.$$

Consider partial derivatives of these functions

$$\frac{\partial g_1(t_1,t_2)}{\partial t_1} = \frac{t_2}{t_1^2}, \frac{\partial g_1(t_1,t_2)}{\partial t_2} = -\frac{1}{t_1},$$

$$\frac{\partial g_2(t_1,t_2)}{\partial t_1} = \frac{t_1(t_1-2t_2)}{(t_1-t_2)^2}, \frac{\partial g_2(t_1,t_2)}{\partial t_2} = \frac{t_1^2}{(t_1-t_2)^2}.$$

Also taking into consideration that $\mu_1 = E(\bar{X}) = E(X) = mp$ and $\mu_2 = E(S^2) = Var(X) = mp(1-p),$ we can write that

$$g_1(\mu_1, \mu_2) = \frac{mp - mp(1-p)}{mp} = p,$$

$$g_2(\mu_1, \mu_2) = \frac{(mp)^2}{mp - mp(1-p)} = m,$$

$$\frac{\partial g_1(\mu_1, \mu_2)}{\partial t_1} = \frac{mp(1-p)}{(mp)^2} = \frac{1-p}{mp},$$

$$\frac{\partial g_1(\mu_1, \mu_2)}{\partial t_2} = -\frac{1}{mp},$$

$$\frac{\partial g_2(\mu_1, \mu_2)}{\partial t_1} = \frac{mp\left(mp - 2mp(1-p)\right)}{\left(mp - mp(1-p)\right)^2} = \frac{2p-1}{p^2},$$

$$\frac{\partial g_2(\mu_1,\mu_2)}{\partial t_2} = -\frac{(mp)^2}{\big(mp - mp(1-p)\big)^2} = \frac{1}{p^2}.$$

By the delta method, the random vector $\sqrt{n}\begin{pmatrix}\hat{p}_n - p\\ \hat{m}_n - m\end{pmatrix}$ is asymptotically normal $N(\vec{0},\tilde{\Sigma})$ with zero mean $\vec{0} = \begin{pmatrix}0\\0\end{pmatrix}$ and covariance matrix $\Sigma = B\tilde{\Sigma}B'$, where

$$B = \begin{pmatrix}\dfrac{1-p}{mp} & -\dfrac{1}{mp}\\ \dfrac{2p-1}{p^2} & \dfrac{1}{p^2}\end{pmatrix},$$

and $\tilde{\Sigma}$ as presented in Lemma 3.1. Simple multiplication of matrices gives the expressions for the matrix Σ as stated in the Theorem.

Remark. To provide more details in the proof of the theorem above, we note that the expansion of functions $\hat{p}_n$ and $\hat{m}_n$ into two dimensional Taylor series by the powers of $\bar{X} - mp$ and $S^2 - mp(1-p),$ keeping only linear terms takes the form

$$\hat{p}_n = p + \frac{1-p}{mp}(\bar{X} - mp) - \frac{1}{mp}\Big(S^2 - mp(1-p)\Big) + R_1,$$

$$\hat{m}_n = m + \frac{2p-1}{p^2}(\bar{X} - mp) + \frac{1}{p^2}\Big(S^2 - mp(1-p)\Big) + R_2,$$

or, equivalently

$$\sqrt{n}\begin{pmatrix}\hat{p}_n - p\\ \hat{m}_n - m\end{pmatrix} = \sqrt{n}\begin{pmatrix}\dfrac{1-p}{mp}(\bar{X} - mp) - \dfrac{1}{mp}\Big(S^2 - mp(1-p)\Big)\\ \dfrac{2p-1}{p^2}(\bar{X} - mp) + \dfrac{1}{p^2}\Big(S^2 - mp(1-p)\Big)\end{pmatrix} + \sqrt{n}\begin{pmatrix}R_1\\ R_2\end{pmatrix}.$$

3.2. Modified estimators

As we already mentioned, the method of moments estimators $\hat{p}_n$ and $\hat{m}_n$ do not have mean values and variances. Therefore, it is incorrect to interpret the parameters of its asymptotic normal approximation as accuracy characteristics (such as bias and variance) of these estimators. But it is possible to have such interpretation for the modified estimators

$$\tilde{p}_n = \frac{\bar{X} - S^2}{\bar{X} + \varepsilon_n}, \quad \tilde{m}_m = \frac{\bar{X}^2}{\bar{X} - S^2 + \varepsilon_n},$$

where ε_n is infinitely small sequence of numbers, that is, $\varepsilon_n \to 0$ sufficiently fast as $n \to \infty$.

According to the Theorem 3.2 below, it is possible to disregard the values of ε_n and interpret the parameters of asymptotic normality as accuracy characteristics of the method of moments estimators.

Theorem 3.2 *Let* $\{\varepsilon_n, n \geq 1\}$ *be a sequence of numbers such that* $\sqrt{n}\varepsilon_n \to 0$ *as* $n \to \infty$. *Then the distribution of the random vector based on the corrected estimators* $\sqrt{n}\begin{pmatrix} \tilde{p}_n - p \\ \tilde{m}_n - m \end{pmatrix}$ *is asymptotically normal* $N(\vec{0}, \Sigma)$ *with the same parameters as the random vector based on the method of moments estimators presented in Theorem 3.1.*

Proof: According to the delta method, we consider the expansion of the functions $\tilde{p}_n$ and $\tilde{m}_n$ into two-dimensional Taylor series, keeping only the linear terms.

Note that

$$\tilde{p}_n = \frac{\bar{X} - S^2}{\bar{X} + \varepsilon_n}, \quad \text{that is,} \quad g_1(t_1, t_2) = \frac{t_1 - t_2}{t_1 + \varepsilon_n},$$

$$\tilde{m}_n = \frac{\bar{X}^2}{\bar{X} + \varepsilon_n - S^2} \quad \text{that is,} \quad g_2(t_1, t_2) = \frac{t_1^2}{t_1 - t_2 + \varepsilon_n}.$$

Partial derivatives:

$$\frac{\partial g_1(t_1, t_2)}{\partial t_1} = \frac{t_2 + \varepsilon_n}{(t_1 + \varepsilon_n)^2} \quad \frac{\partial g_1(t_1, t_2)}{\partial t_2} = -\frac{1}{t_1 + \varepsilon_n},$$

$$\frac{\partial g_2(t_1, t_2)}{\partial t_1} = \frac{t_1(t_1 - 2t_2)}{(t_1 - t_2 + \varepsilon_n)^2} \quad \frac{\partial g_2(t_1, t_2)}{\partial t_2} = \frac{t_1^2}{(t_1 - t_2 + \varepsilon_n)^2}.$$

Let $\alpha_n = \frac{\varepsilon_n}{mp}$, then $\alpha_n \to 0$. Taking into consideration that $\mu_1 = E(\bar{X}) = mp$ and $\mu_2 = Var(X) = mp(1-p)$, we can write that

$$g_1(\mu_1, \mu_2) = \frac{mp - mp(1-p)}{mp + \varepsilon_n} = \frac{p}{1 + \alpha_n},$$

$$g_2(\mu_1, \mu_2) = \frac{(mp)^2}{mp - mp(1-p) + \varepsilon_n} = \frac{m}{1 + \alpha_n / p},$$

$$\frac{\partial g_1(\mu_1, \mu_2)}{\partial t_1} = \frac{mp(1-p)}{(mp + \varepsilon_n)^2} = \frac{1 - p + \alpha_n}{mp(1 + \alpha_n)^2},$$

$$\frac{\partial g_1(\mu_1, \mu_2)}{\partial t_2} = -\frac{1}{mp + \varepsilon_n} = -\frac{1}{mp(1 + \alpha_n)},$$

$$\frac{\partial g_2(\mu_1, \mu_2)}{\partial t_1} = \frac{mp\left(mp - 2mp(1-p)\right)}{(mp - mp(1-p) + \varepsilon_n)^2} = \frac{2p - 1}{p^2(1 + \alpha_n / p)},$$

$$\frac{\partial g_2(\mu_1, \mu_2)}{\partial t_2} = \frac{(mp)^2}{(mp - mp(1-p) + \varepsilon_n)^2} = \frac{1}{p^2(1 + \alpha_n / p)^2}.$$

By the delta method, the random vector $\sqrt{n}\begin{pmatrix} \tilde{p}_n - \dfrac{p}{1+\alpha_n} \\ \tilde{m}_n - \dfrac{m}{1+\alpha_n / p} \end{pmatrix}$ converges in distribution to the bivariate normal $N(\vec{0}, \Sigma)$ with zero mean $\vec{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\Sigma = \tilde{B}_n \Sigma \tilde{B}_n'$, where

$$\tilde{B}_n = \begin{pmatrix} \dfrac{1-p+\alpha_n}{mp(1+\alpha_n)^2} & -\dfrac{1}{mp(1+\alpha_n)} \\ \dfrac{2p-1}{p^2(1+\alpha_n/p)} & \dfrac{1}{p^2(1+\alpha_n/p)^2} \end{pmatrix},$$

and Σ is as stated in the Theorem (matrix $\tilde{\Sigma}$ is the covariance matrix for $\bar{X}$ and S^2, it is derived in Lemma 3.1).

To be more precise,

$$\begin{pmatrix} g_1\left(\bar{X}_n, S_n^2\right) - g_1\left(\mu_1, \mu_2\right) \\ g_2\left(\bar{X}_n, S_n^2\right) - g_2\left(\mu_1, \mu_2\right) \end{pmatrix} = \begin{pmatrix} \tilde{p}_n - \dfrac{p}{1+\alpha_n} \\ \tilde{m}_n - \dfrac{m}{1+\alpha_n / p} \end{pmatrix}$$

$$= \begin{pmatrix} (\bar{X} - mp)\dfrac{1-p+\alpha_n}{mp(1+\alpha_n)^2} - \left(S^2 - mp(1-p)\right)\dfrac{1}{mp(1+\alpha_n)} \\ (\bar{X} - mp)\dfrac{2p-1}{p^2(1+\alpha_n/p)} + \left(S^2 - mp(1-p)\right)\dfrac{1}{p^2(1+\alpha_n/p)^2} \end{pmatrix} + \begin{pmatrix} R_{1n} \\ R_{2n} \end{pmatrix}.$$

It is important to note that the reminder term $\sqrt{n}\begin{pmatrix} R_{1n} \\ R_{2n} \end{pmatrix} \to \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ in probability as $n \to \infty$. The last expression can be rewritten as

$$\begin{pmatrix} \tilde{p}_n - p \\ \tilde{m}_n - m \end{pmatrix} = \begin{pmatrix} (\bar{X} - p)\dfrac{1-p+\alpha_n}{mp(1+\alpha_n)^2} - (S^2 - mp)\dfrac{1}{mp(1+\alpha_n)} \\ (\bar{X} - p)\dfrac{2p-1}{p^2(1+\alpha_n/p)} + (S^2 - mp)\dfrac{1}{p^2(1+\alpha_n/p)^2} \end{pmatrix} + \begin{pmatrix} R_{1n} - p + \dfrac{p}{1+\alpha_n} \\ R_{2n} - m + \dfrac{m}{1+\alpha_n/p} \end{pmatrix}.$$

Consider a matrix

$$A_n = \tilde{B} - B = \begin{pmatrix} \dfrac{1-p+\alpha_n}{mp(1+\alpha_n)^2} - \dfrac{1-p}{mp} & -\dfrac{1}{mp(1+\alpha_n)} + \dfrac{1}{mp} \\ \dfrac{2p-1}{p^2(1+\alpha_n/p)} - \dfrac{2p-1}{p^2} & \dfrac{1}{p^2(1+\alpha_n/p)^2} - \dfrac{1}{p^2} \end{pmatrix}.$$

Then the expression above can be rewritten

$$\begin{pmatrix} \tilde{p}_n - p \\ \tilde{m}_n - m \end{pmatrix} = B\begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix} + A_n\begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix} + \begin{pmatrix} R_{1n} + \left(p - \dfrac{p}{1+\alpha_n}\right) \\ R_{2n} + \left(m - \dfrac{m}{1+\alpha_n/p}\right) \end{pmatrix}.$$

Note that the numerical matrix $A_n \to \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ because $\alpha_n \to 0$ as $n \to \infty$. Hence

$$\sqrt{n} A_n \begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix} = A_n \sqrt{n} \begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix} \to \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

in probability because $\sqrt{n} \begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix}$ is bounded in probability.

Denote the new reminder term

$$\begin{pmatrix} \tilde{R}_{1n} \\ \tilde{R}_{2n} \end{pmatrix} = \begin{pmatrix} R_{1n} + \left(p - \dfrac{p}{1+\alpha_n} \right) \\ R_{2n} + \left(m - \dfrac{m}{1+\alpha_n / p} \right) \end{pmatrix} + A_n \begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix},$$

then we can write

$$\begin{pmatrix} \tilde{p}_n - p \\ \tilde{m}_n - m \end{pmatrix} = B \begin{pmatrix} \bar{X} - mp \\ S^2 - mp(1-p) \end{pmatrix} + \begin{pmatrix} \tilde{R}_{1n} \\ \tilde{R}_{2n} \end{pmatrix},$$

where the remainder term $\begin{pmatrix} \tilde{R}_{1n} \\ \tilde{R}_{2n} \end{pmatrix}$ has the property that $\sqrt{n} \begin{pmatrix} \tilde{R}_{1n} \\ \tilde{R}_{2n} \end{pmatrix} \to \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ according to our remarks above and the assumptions of the theorem.

4. Confidence Regions and Hypotheses Testing for Parameters of the Binomial Distribution

The asymptotic results presented in the previous sections provide us a way to statistical inference for the parameters m and p of the binomial distribution. The main tool here is the observation that according to Lehmann (2004) Theorem 5.4.2 (ii), if random vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ is normally distributed with the mean vector $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)$ and covariance matrix Σ, then the quadratic form $(\vec{X} - \vec{\xi})' \Sigma^{-1} (\vec{X} - \vec{\xi})$ is distributed as χ^2 with k degrees of freedom.

In case of the binomial distribution, there are two parameters, so $k = 2$ for both method of moments and modified estimates, the covariance matrix $\Sigma = \begin{pmatrix} \sigma_p^2 & \rho \sigma_p \sigma_m \\ \rho \sigma_p \sigma_m & \sigma_m^2 \end{pmatrix}$, and mean vector $\vec{\xi} = \begin{pmatrix} p \\ m \end{pmatrix}$. It is well known that the inverse matrix $\Sigma^{-1} = \dfrac{1}{1-\rho^2} \begin{pmatrix} \dfrac{1}{\sigma_p^2} & -\dfrac{\rho}{\sigma_p \sigma_m} \\ -\dfrac{\rho}{\sigma_p \sigma_m} & \dfrac{1}{\sigma_m^2} \end{pmatrix}$. Substituting the values of σ_p, σ_m and ρ, presented in Theorem 3.1, we obtain that

$$\Sigma^{-1} = \left(\frac{p+a}{p} \right) \begin{pmatrix} \dfrac{m}{(1-p)(p+a)} & -\dfrac{p}{(1-p)(+ap)} \\ -\dfrac{p}{(1-p)(p+a)} & \dfrac{p^2}{m(1-p)a} \end{pmatrix} = \frac{1}{p(1-p)} \begin{pmatrix} m & -p \\ -p & \dfrac{p^2(p+a)}{ma} \end{pmatrix},$$

where $a = 2(1-p)(m-1)$. Therefore, we can state that the statistic

$$X_2^2(p,m) = \frac{1}{p(1-p)}(\hat{p}-p, \hat{m}-m)\begin{pmatrix} m & -p \\ -p & \frac{p^2(p+a)}{ma} \end{pmatrix}\begin{pmatrix} \hat{p}-p \\ \hat{m}-m \end{pmatrix},$$

$$= \frac{1}{p(1-p)}\left(m(\hat{p}-p)^2 - 2p(\hat{m}-m)(\hat{p}-p) + \frac{p^2(p+a)}{ma}(\hat{m}-m)^2 \right),$$

is asymptotically distributed as χ^2 with 2 degrees of freedom.

For hypothesis testing for $H_0 : p = p_0$ and $m = m_0$ with significance level α, we use the statistic $X_2^2(p_0, m_0)$ and compare it with the quantile $\chi_2^2(\alpha)$ of the χ^2 distribution with 2 degrees of freedom.

For the construction of the $100(1-\alpha)\%$ confidence region for both parameters p and m we consider the following ellipse $\left\{(p,m) \mid X_2^2(p,m) \le \chi_2^2(\alpha)\right\}$.

Exactly the same conclusions are true if we consider the modified estimators $\tilde{p}$ and $\tilde{m}$ instead of method of moments estimators $\hat{p}$ and $\hat{m}$. Simply change symbol in the formulae above.

5. Conclusions and Future Plans

In this paper, we present a sufficiently complete review of the literature pertaining to the problem of estimating parameters for the binomial distribution. As the main theoretical contribution, we derive formulae for estimators of the binomial distribution by the method of moments and prove their joint asymptotic normality in Theorem 3.1. Also, modified estimators are introduced, and their joint asymptotic normality is shown in Theorem 3.2. We also provided the guidelines of constructing hypothesis testing and confidence intervals for the binomial parameters.

There will be two papers more to come on this subject. In the second paper, we will compare the derived asymptotic with the true probabilistic characteristics of the estimators by the method of statistical simulations. In the third paper, we will discuss the real life problem of an application of our techniques to the estimation of parameters m and p by observations of the number of replies in a response for a nerve stimulation in an experiment on the neuromuscular junction (synapse) (m is the number of synaptic vesicles with the neurotransmitter acetylcholine in a nerve ending and p is the probability of acetylcholine release by a synaptic vesicle).

Acknowledgements

The last listed author's research was partially supported by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project 1.13556.2019/13.1.

References

Binet FE. The fitting of the positive binomial distribution when both parameters are estimated from the sample. Ann. Eugenics. 1952; 17(1): 117-119.

Blumenthal S, Dahiya RC. Estimating the binomial parameter n. J Am Stat Assoc. 1981; 76(376): 903-909.

Carroll RJ, Lombard F. A note on N estimators for the binomial distribution. J Am Stat Assoc. 1985; 80(390): 423-426.

Cramér H. Mathematical methods of statistics. Princeton: Princeton University Press; 2016.

DasGupta A, Rubin H. Estimation of binomial parameters when both n, p are unknown. J Stat Plan Infer. 2005; 130(1-2): 391-404.

Deriggi DF. Unimodality of likelihood functions for the binomial distribution. J Am Stat Assoc. 1983; 78(381): 181-183.

Draper N, Guttman I. Bayesian estimation of the binomial parameter. Technometrics. 1971; 13(3): 667-673.

Feldman D, Fox M. Estimation of the parameter n in the binomial distribution. J Am Stat Assoc. 1968; 63(321): 150-158.

Fisher RA. The negative binomial distribution. Ann Eugenic. 1941; 11(1): 182-187.

Haldane JBS. The fitting of binomial distributions. Ann Eugenic. 1941; 11(1): 179-181.

Hall P. On the erratic behavior of estimators of N in the binomial N, p distribution. J Am Stat Assoc. 1994; 89(425): 344-352.

Hammersley JM. On estimating restricted parameters. J R Stat Soc B. 1950; 12(2): 192-240.

Hoel PG. Discriminating between binomial distribution. Ann Math Stat. 1947; 18(4): 556-564.

Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1): 55-67.

Jeffreys H. The theory of probability. New York: Oxford University Press; 1998.

Lehmann EL. Elements of large-sample theory. New York: Springer-Verlag; 2004.

Nicholls JG, Martin AR, Wallace BG, Fuchs PA. From neuron to brain. Massachusetts: Sunderland; 2001.

Olkin I, Petkau AJ, Zidek JV. A comparison of n estimators for the binomial distribution. J Am Stat Assoc. 1981; 76(375): 637-642.

Raftery AE. Inference for the binomial N parameter: a hierarchical Bayes approach. Biometrika. 1988; 75(2): 223-228.

Saad S. Asymptotic Ananlysis of method of moments estimators of parameters p and m for the binomial distribution. PhD Thesis, University of Regina; 2019a.

Saad S. Joint normality of the sample mean and sample variance. J Prob Stat Sci. 2019b; 17(2): 223-226.

Skellam JG. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. J R Stat Soc B. 1948; 10: 257-261.

van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 1998.