Available online at www.garmian.edu.krd

Passer Journal

Passer 2 (2020) 63- 68

http://passer.garmian.edu.krd/

Beta–Binomial and Overdispersion with Exchange of the Sample Size over the Probability Interval [0, 1] with Applications

Hanaw Ahmed Amin[1 *], Rando Rasul Qadir[2]

[1] *Department of Mathematics, College of Science, University of Sulaimani, Sulaimani, Kurdistan Region, Iraq*
[2]*Department of Mathematics, College of Basic Education, University of Sulaimani, P.O. Box: 46, Sulaimani, Kurdistan Region, Iraq*

Received 20 August 2020; revised 08 October 2020;
accepted 08 October 2020; available online 22 October 2020

doi:10.24271/psr.13

ABSTRACT

The beta-binomial model that is generated by a simple mixture model has been commonly applied in the health, physical, and social sciences. In clinical and public health, overdispersion occurs due to biological variation between the subjects of interest. Both the binomial and beta-binomial models are applied to different problems occurring in rational test theory. In this study, we focused on modeling overdispersion for binomial distribution. The main aim was to show a complete and extensive understanding of the beta-binomial model and updated form by broaden its practical applications in the field of breast cancer with hormone medication. It is observed in different independent Bernoulli trials yes/no $(x_i = 1, 0)$ experiments with success probabilities $0< p_i < 1$ and compare the model in a sequence of n_i. The performance of the maximum likelihood estimates technique that is used in moderate and small samples n_i by a Newton-Raphson iterative method using Matlab package. We have found that using hormones for other treatments have complication leading to breast cancer. We took 20 investigational testers in Hiwa Hospital for cancer treatment in Sulaymaniyah province, with proportion p_i is varying from 9.7% to 50 %. In addition, we concluded that the beta-binomial theory is a good alternative of binomial model. This is due to the fact that the beta-binomial model has provided a robust estimate for events from heterogeneous binomial studies.

Keywords: Beta-Binomial, Overdispersion, Maximum Likelihood Estimation, Binary Data, Breast Cancer, Tarone'Z Test.

1. Introduction

In statistics, overdispersion is the occurrence of bigger changeability in a data set rather than would be predictable in a statistical model. When the value of the determined variance is higher than its value in a theoretical model, one can say overdispersion will be observed [1]. On the other hand, under dispersion is an indication of less variation in the data than assumed. Overdispersion is an ordinary feature in applied data analysis; this is due to the fact that populations are frequently heterogeneous (non-uniform) contrary to the assumptions implied within broadly used simple parametric models [1]. Pearson introduced the beta-binomial model in 1925 and then more formally described by Skellam in1948 which is a general method for obviously explanation for the overdispersion [2]. For example, this model has several applications in different areas, such as explained by Chatfield and Goodhardt in 1976 for buying performance of the user, and Gange in1996, who considered the impact of policy changes on suitability of hospital admissions. In addition, in 1977, Aeschbacher showed that a beta-binomial distribution offers a better application than the typical distribution in biological tests involving mice when the data used were based on a large number of death counts. Single parameter distributions, such as Poisson and binomial, suggest that the variance is determined by the mean value [3]. In many cases and especially in analysis of biological data, the mean-variance relationship fails generally due to existence of overdispersion, where the data have a higher variance than anticipated under the simple model of Cox (1983), Hinde & Demetrio (1998) [4]. A beta-binomial distribution is a combination of binomial and beta distribution and it is one of the simplest Bayesian models. A distribution is beta-binomial with the probability of success p, in a binomial distribution has a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ [4], [5].

In this paper, we are going to examine the beta-binomial distribution model for a 20 samples of cancer cases for various number of patients who are tested positive for breast cancer. The causes of cancer cases are correlated to the history of patients who

* Corresponding author
E-mail address: hanaw.ammin@univsul.edu.iq (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

took hormones for other treatment, and having family history, genetics, obesity, age, menstruation, residence, smoking, alcohol, etc.

2. The Review of Beta-Binomial Distribution

The investigation of proportions has been communicated about commonly from a varied collection of views. A representative idea is whether the data follows a binomial or multinomial distribution. In some situation, a couple authors, Kleinman and Lee, have considered that as far as they can tell information which gives off an impress of presence binomial proportions now and again show difference which outcomes in further projecting variance than would be normal in the binomial distribution conditions [2], [5], [6]. The beta-binomial model, after associated with numerous examinations, can be realized as developing as of two-organize methods. If $\boldsymbol{x}$ is a random variable, which is distributed with binomial distribution, $\boldsymbol{B}(\boldsymbol{x};\boldsymbol{n},\boldsymbol{p})$, then the probability mass function of x is given by:

$$B(x;n,p)=\binom{n}{x}p^{x}(1-p)^{n-x},\ 0\leq p\leq 1. \qquad (1)$$

Where n is the sample size of the data and p is the probability of success distinct x. The mean and variance of the binomial distribution are np and $np(1-p)$, respectively. Supposing that the probability of success p is distributed with Beta distribution, $Bet(p;\alpha,\beta)$, then the probability density function of p is defined as:

$$Bet(p;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}. \qquad (2)$$

For some $\alpha,\beta>0$ are two positive parameters and Γ is the gamma function in the domain $[0,1]$, then $\mu=\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}=\gamma\mu(1-\mu)$ are the mean and the variance of beta distribution, respectively, where $\gamma=\frac{1}{\alpha+\beta+1}$. The beta-binomial distribution is a mix of (1) and (2), which is represented by $BB(x,n;\alpha,\beta)$ which can be find in [7]. In other words, if x is a random variable where distributed with this combined distribution, then the probability mass function of x is:

$$BB(x,n;\alpha,\beta)=\binom{n}{x}\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}p^{\alpha-1}(1-p)^{\beta-1}. \quad (3)$$

Where, $x=0,1,\dots,n$ and $\alpha,\beta>0$. It can be realized that n is the sample size of all individuals and x is the total number of concerned (success) in the data. It is recognized that both mean and variance are $\frac{n\alpha}{\alpha+\beta}=n\mu$ and $n\mu(1-\mu)+\gamma\mu(1-\mu)n(n-1)$ of the beta-binomial distribution, for $\gamma=\frac{1}{\alpha+\beta+1}$ [3], [8].

Now, supposing that we have k different samples, $x_i,(i=1,\dots,k)$ is the number of success in the i^{th} sample and $n_i,(i=1,\dots,k)$ in the experiment. If p_i is denote the proportion $\frac{x_i}{n_i}$, for $i=1,\dots,k$, then there will be two stage model:

$$x_i\sim B(x_i;n_i,p_i),\ 0\leq p_i\leq 1. \qquad (4)$$

$$p_i\sim Bet(p_i;\mu,\theta)\ i.i.d. \qquad (5)$$

Where $\mu=\frac{\alpha}{\alpha+\beta}$ is the mean and $\theta=\frac{1}{\alpha+\beta}$ is an amount of variance of beta distribution, then we can certainly estimate the mean of x_i is μ and its variance is $\frac{\mu(1-\mu)}{n_i}\left(\frac{n_i\theta+1}{1+\theta}\right)$, it can be seen that $\left(\frac{n_i\theta+1}{1+\theta}\right)$ is the amount multiplier of the variance of binomial distribution. In this study, the means and variance of x_i in terms of α and β are estimated also, we compare the results to the estimations in terms of μ and θ in [2], [9], [10]. The beta-binomial probability distribution with different values of parameters α,β and constant n is plotted in Figure 1, the cumulative distribution is plotted in Figure 2. Special cases of the binomial distribution include the beta-binomial distribution $n=10$, $\beta=400$ and $\alpha=600$, respectively.

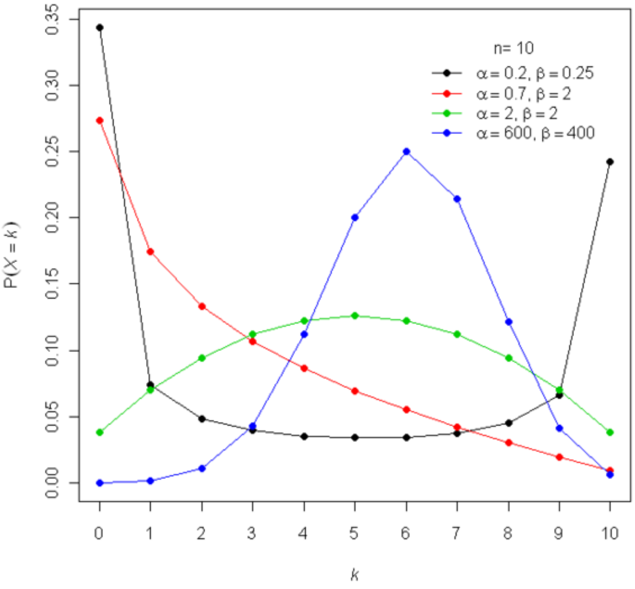

Figure 1: The probability function of beta-binomial random variable with some specific values of, and of(n),$(\alpha)$and (β)

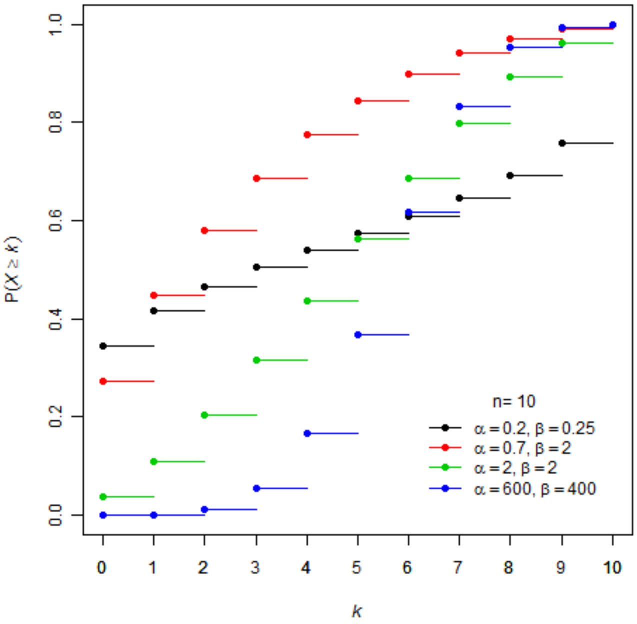

Figure 2: The cumulative distribution function of beta- binomial with some specific values of (n,α,β)

2.1 Estimation of Parameters

The moment and maximum likelihood estimations are two techniques for estimating the parameters α and β (or μ and θ). In the first subsection, we review the moment estimation and in the second subsections the maximum likelihood estimation and derivations of the two methods can found in [9].

2.1.1 Moment Estimation of "μ" and "θ"

The techniques of estimating beta-binomial distribution considered by many authors for example Kleinman is one of them that considered beta-binomial distribution, he provided an evaluation for the mean by using moments, he determined a greatest sufficiency results, since this technique is certainly evaluate the parameters, compared by the other techniques [4], [5], [11]. The steps of Kleinman can be view as follows [7], [9], [12]:

$$\text{If } \hat{p} = \sum_{i=1}^{k} \frac{w_i p_i}{w},\ w = \sum_{i=1}^{k} w_i \text{ and } w_i = \frac{n_i}{1+\gamma(n_i-1)}. \tag{6}$$

Suppose the least square approximation$S = \sum_{i=1}^{k} w_i\, (p_i - \hat{p})^2$. Consider, $\hat{p}$ and S are equal to their expectation, then we get the estimation for μ and γ:

$$\hat{\mu} = \hat{p}, \text{ and } \hat{\gamma} = \frac{S - \hat{p}\hat{q}\left[\sum_{i=1}^{k} \frac{w_i}{n_i}\left(1-\frac{w_i}{w}\right)\right]}{\hat{p}\hat{q}\left[\sum_{i=1}^{k} w_i\left(1-\frac{w_i}{w}\right) - \sum_{i=1}^{k} \frac{w_i}{n_i}\left(1-\frac{w_i}{w}\right)\right]} \tag{7}$$

Where $\hat{q} = 1 - \hat{p}$. Hence, we can find the value of θ, by using the relation $\theta = \frac{\gamma}{1-\gamma}$.

The moment estimation depends on $\{w_i\}$, the choice of the weights. This is a notable outcome as $\{w_i\}$ are been reciprocally corresponding to the variance of p_i, at that point $\hat{p}$ has the minimum variance among entire linear unbiased estimations of μ. This characteristic drove Kleinman to think about the accompanying weights: [13]

$$\mathrm{w_i} = \frac{\mathrm{n_i}}{1+\gamma(\mathrm{n_i}-1)} \tag{8}$$

The complexity of the estimations of Kleinman began when equation (8) is a function of the variable γ, which we want to estimate it. So, Kleinman puts the values of the weighting $\mathrm{w_i} = \mathrm{n_i}$ or $\mathrm{w_i} = 1$, which is a result of putting $\gamma = 1$ or $\gamma = 0$. Chuang-Stein suggested refining Kleinman's empirical weighting form by continuing more steps [6]. Especially, an estimate for γ can be deduced using the parameter $\hat{p}$, the corresponding weights, along with equation (8). This new γ estimation can be utilized to renew w_i resulting other estimations for both μ and γ. This procedure can be repeated many times until the difference between estimations for both of μ and γ will be smaller than the order of 10^{-6}.

2.1.2 Maximum Likelihood Estimation of the Parameters "μ" and "θ"

If x is the number of responses of a sample with size n of subjects that responding to the particular dosage, then the proportion $\frac{x}{n}$ is a significant form of biological, physical, chemical and real-life discrete data. Binomial distribution is one of those distributions that can be used to estimate the mean and variance of the data, but in some cases the variance of the data is a large value, so there is overdispersion, in such a case beta-binomial is a common distribution that can be used to reduce the variance and compromise the over dispersion [1], [12], [14]. The maximum likelihood technique is a way to estimate the values of parameters of α and β consequently, μ and θ for the data, to give a best approximation with a small variance [7]. Suppose that we have k different sample sizes $n_1, n_2, \dots, n_k$ with k different numbers $x_1, x_2, \dots, x_k$ that responding response, then the equation for the maximum likelihood of the beta-binomial distribution according to the parameters α and β is:

$$L(\alpha,\beta) = \prod_{i=0}^{k} \binom{n_i}{x_i} \frac{Beta(\alpha+x_i,\beta+n_i-x_i)}{Beta(\alpha,\beta)}. \tag{9}$$

Such that $Beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function. The best approximate values are the solution of the log likelihood method, that is we can use the logarithm function for equation (9) to obtain a best approximation for the data, and we get the log likelihood equation:

$$c - \sum_{i=1}^{k} n_i \log\bigl(Beta(\alpha,\beta)\bigr) + \sum_{i=0}^{k} n_i \log\bigl(Beta(\alpha + x_i, \beta + n_i - x_i)\bigr). \tag{10}$$

Where c is a constant, then we differentiate (10) with respect to α and β, to get a best solution for fitting the data using the log likelihood equation of Beta-Binomial distribution [7], then we get the following non-linear system equation for estimating α and $\beta$is:

$$\frac{\partial \log L(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{k} \Delta_1(\alpha, x_i) - \sum_{i=1}^{k} \Delta_1(\alpha+\beta, n_i) = 0. \tag{11}$$

$$\frac{\partial \log L(\alpha,\beta)}{\partial \beta} = \sum_{i=1}^{k} \Delta_1(\beta, n_i - x_i) - \sum_{i=1}^{k} \Delta_1(\alpha+\beta, n_i) = 0. \tag{12}$$

The equations (11) and (12) are two complicated systems on non-linear equation, where Δ_1 is a series function of two variables m and n defined as follows [7]:

$$\Delta_1(m,n) = \frac{1}{m+n-1} + \frac{1}{m+n-2} + \cdots + \frac{1}{m}.$$

The log likelihood equation is also can be used to find the covariance of the data. This means, if we differentiate the log likelihood equation two times with respect to α and β, then we get the Hessian matrix, which can be used to deduce the covariance, or standard error, of data, that is

$$H(\alpha,\beta) = \begin{bmatrix} \frac{\partial^2 \log L(\alpha,\beta)}{\partial \alpha^2} & \frac{\partial^2 \log L(\alpha,\beta)}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \log L(\alpha,\beta)}{\partial \alpha \partial \beta} & \frac{\partial^2 \log L(\alpha,\beta)}{\partial \beta^2} \end{bmatrix}$$

Where,

$$\frac{\partial^2 \log L(\alpha,\beta)}{\partial \alpha^2} = -\sum_{i=1}^{k} \Delta_2(\alpha, x_i) + \sum_{i=1}^{k} \Delta_2(\alpha+\beta, n_i),$$

$$\frac{\partial^2 \log L(\alpha,\beta)}{\partial \beta^2} = -\sum_{i=1}^{k} \Delta_2(\beta, n_i - x_i) + \sum_{i=1}^{k} \Delta_2(\alpha+\beta, n_i),$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta} = \sum_{i=1}^{k} \Delta_2(\alpha+\beta, n_i),$$

where, $\Delta_2(m,n)$ is a function of two variable m and n, which is defined by

$$\Delta_2(m,n) = \frac{1}{(m+n-1)^2} + \frac{1}{(m+n-2)^2} + \cdots + \frac{1}{m^2}$$

The inverse of the matrix $H(\alpha,\beta)$, is the covariance matrix and the standard error of α and β. Hence, can estimate α and β by solving the equations (11) and (12), consequently we can find the values of μ and θ from their relation with α and β. In another word, it can be directly used to estimate μ and θ for the two-stage model (4) and (5) as follows:

Suppose that $f_x(x = 0.1.\dots.k)$ is the determined frequencies of k trial, subsequently the maximum likelihood equation converted to [7]

$$L(\alpha,\beta) = \prod_{i=0}^{k} [BB(x,n;\alpha,\beta)]^{f_{x_i}}$$

If $S_i = \sum_{x=0}^{i} f_x$, then $S_k = \sum_{i=1}^{k} n_i = n$, where n is the total sample size of all individual trials combined, then the log likelihood equation for the model (4) and (5) has the form:

$$c - S_n \sum_{i=1}^{n-1}[\log(1+i\theta)] + \sum_{i=0}^{n-1}[(S_n - S_i)\log(\mu+i\theta) + S_{n-1-i}\log(1-\mu+i\theta)]. \quad (13)$$

As in similar way of estimation of a α and β, differentiating (13) with respect to μ and θ, then we get the maximum log likelihood equation, which is a system of non-linear equation functions of the form:

$$\frac{\partial \log L(\mu,\theta)}{\partial \mu} = \sum_{i=0}^{k-1}\left[\frac{S_k - S_i}{\mu+i\theta} - \frac{S_{k-1-i}}{1-\mu+i\theta}\right] = 0. \quad (14)$$

$$\frac{\partial \log L(\mu,\theta)}{\partial \theta} = \sum_{i=0}^{k-1} i\left[\frac{S_k - S_i}{\mu+i\theta} - \frac{S_{k-1-i}}{1-\mu+i\theta} - \frac{S_{k-1-i}}{1+i\theta}\right] = 0. \quad (15)$$

By solving equations (14) and (15), we get a direct estimation of μ and θ. Also, differentiating (13) partially with respect to μ and θ we get the Hessian matrix, which is the inverse of the covariance matrix for data by using the model (4) and (5). Once, we obtain the maximum log likelihood for the data by using any of the above method, we need to obtain the likelihood ratio, which requires maximum log likelihood equation of Binomial distribution [7], this is an equation of:

$$\log L = \sum_{i=1}^{k}\left[\log\binom{n_i}{x_i} + x_i \log p_i + (n_i - x_i)\log(1-p_i)\right]. \quad (16)$$

Then the likelihood ratio is defined as:

$$\chi^2 = 2(L_{BB} - L_B). \quad (17)$$

Where L_B and L_{BB} are the log likelihood of Binomial and Beta-Binomial distributions respectively [12]. A non-linear equation system of (11) and (12), which there solution for α and β is complicated, for this reason we have to use Newton-Raphson technique for resolving these equations, again using Newton-Raphson method requires mathematical software, many authors uses some packages of R program for solving the equations (11) and (12) or equivalently (14) and (15), for example [2], [3], [14]. In this study, we wrote a MATLAB program for solving these equations, with computing the measurements of $\alpha, \beta, \mu, \gamma$ and θ.

2.2 Test for Overdispersion

Before assuming the Beta-Binomial model for analyzing a set of data, it should be tested to find out a problem of overdispersion to the degree where the beta- binomial model would be a better and more acceptable than the simple Binomial model [3], [5], [12]. Since the value of (p)can be estimated, it can be directly tested whether (p) is significantly greater than zero. However, this test has less sensitivity to detect departure from Binomial. This is due to the fact that the boundary problems will appear as we check whether a positive-valued parameter is larger than zero.

There are several ways to examine overdispersion, knowing that [3]

$$E(p_i) = \mu = \frac{\alpha}{\alpha+\beta} ,$$

$$\mathrm{V(p_i)} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = \gamma\mu(1-\mu)$$

$$\gamma = \frac{1}{(1+\alpha+\beta)}$$

If we have a way to estimate the value γ, then we can roughly know whether γ is zero. There will be no significant overdispersion as the value of γ is close to zero, and thus the binomial model will completely define the data.

2.2.1 Likelihood Ratio Test:

Another approach is the likelihood ratio test (LRT) (16). In the null hypothesis, the underlying distribution is referred to as binomial. Whereas in the alternative hypothesis, the distribution is referred to as beta-binomial. The likelihood ratio test statistic is [2], [3], [12].

This $\chi2$ test statistic (17) agrees well with a $\chi2$ distribution that has 1 degree of freedom, which is difference on the number of parameters for each distribution. It should be noted that the same boundary problem applies also for this test.

2.2.2 Tarone's Z statistic:

In order to not face the boundary problem, an alternative statistic called Tarone's Z statistic Tarone (1979) will be employed. This statistic can be used well as a fit test of the Binomial distribution beside the BB distribution [12],

$$Z = \frac{E - \sum_{i=1}^{k} n_i}{\sqrt{2\sum_{i=1}^{k} n_i(n_i-1)}} \quad (18)$$

where

$$\mathrm{E} = \sum_{i=1}^{k} \frac{(x_i - n_i p)^2}{p(1-p)}$$

$$\mathrm{P} = \sum_{i=1}^{k} \frac{x_i}{n_k}$$

The statistic Z has known to have an asymptotic standard which is a normal distribution of binomial distribution under the null hypothesis. [12], [15]

2.3 Applications of Maximum Likelihood of Beta-Binomial Distribution

Here, the maximum log likelihood of beta-binomial distribution (10) will be used for compromise the overdispersion of binomial distribution. At the first example we use the table of Chuang-Stein's article [6], which is a data of 15 studies of patients that stimulated with anti-cancer chemical compound that play the role of inducing cardiac toxicity. In his paper, Chuang-Stein used the moment methods of Kleinman [2] for estimating the parameters μ and θ. In this paper, we use maximum likelihood of beta-binomial distribution for computing α and β directly, then we estimate μ and θ, by which the mean, variance, and covariance for the data can be compute easily. In Table (1) we put the cancer data which is the responding of the patients for the breast cancer with hormone medication of 20 different samples.

Table 1: Cases of Hormone Medication of 20 Different Samples Breast Cancer which are taken in Hiwa Hospital for Cancer Treatment in Sulaymaniyah Province.

Studies	Sample sizes n_i	Number of Patients with Hormone Treatment x_i	Proportion of Hormone Treatment $p_i = \frac{x_i}{n_i}$
1	21	6	28.57 %
2	23	6	26 %
3	19	6	31.57 %
4	20	2	10 %
5	15	3	50 %
6	19	9	47%
7	22	8	36%
8	31	11	35%
9	50	15	30%
10	120	31	25.8%
11	180	18	10%
12	60	16	26.67%
13	200	66	33%
14	70	18	25.7%
15	230	65	28%
16	260	40	15%
17	300	30	10%
18	275	27	9.8%
19	245	24	9.7%
20	40	11	27.5%

Table (3) contains all results by using different methods to evaluate overdispersion. As explained in the former section, $\gamma =$ 0.0194 is the clear sign of existing overdispersion problem. As it is significantly larger than zero, in this case where $(p < 0.05)$, confirming the existence of overdispersion. The value of Tarone's $Z = -40.2322$ statistics; fall in the reject region according to the tabulated $Z_{\frac{\alpha}{2}} = 1.960$. This result indicates that beta- binomial has well-fitting than binomial model. After using the beta- binomial model in Table (2), the summary event rates are $\hat{\mu}$ =1.3033% with a predicted value of standard error 0.01563%. The $\hat{\theta}$ is estimated to be 0.0198 (Table 3), which results an α estimation of 2.99487318 and an β estimate of 47.55704330. Once these parameters are predicted, Equation (3) can be used of beta- binomial model, as a prediction equation for observing new patients and new sample sizes with the probability of 0.05%.

Table 2: Prediction of Proportion Rate

Methods	Estimate	Standard Error	Lower CI 95%	Upper CI 95%
Simple Binomial	4.12%	0.390%	3.374%	4.833%
Beta-Binomial	1.3033%	0.01563%	0.86%	1.742%

Table 3: Estimated Parameters of the Beta-Binomial Distribution

Alpha	2.99487318
Beta	47.55704330
Mean	0.0592
Theta	0.0198
Gamma	0.0194
Variance	0.0227
Tarone's Z	-40.2322

3. Result and Discussion:

In an example regarding the effect of hormone usage in identification in females breast cancer, we ensured 20 investigational testers, with p_i rates varying from 9.7% to 50 %. When beta-binomial model is used, the event rates is varying from $\hat{\mu}$ =1.3033% a standard error 0.01563% are obtained, also beta-binomial model has the values of $\alpha = 2.99487318$ and $\beta = 47.55704330$, the event rate of the binomial model is 4.12% with a standard error of 0.390%, but beta-binomial model can provide a robust estimate for events from heterogeneous binomial studies. Since the important parameters α and β can be estimated and there are k different sample sizes $n_1, n_2, \dots, n_k$ with k different numbers $x_1, x_2, \dots, x_k$. The best convergence between the two-distribution binomial and beta-binomial can be determined using these different samples by substituting the values of α, $\beta$and different n_i into a beta-binomial probability function (3) as shown in Figure 1 and Figure 2.

The results showed that 3% of the breast cancer is due to the effect of the hormones. Furthermore, 97% of breast cancer could be due to family history, genetics, obesity, age, menstruation, residence, smoking, alcohol, etc.

After looking briefly at the binomial, beta, and beta- binomial distributions with their properties and the relation between them, how they act when their parameters are changed. This distribution has been tested for different sample size (n), in the case of overdispersion indicate that:

1- ML estimates of the performance in moderate and small samples indicate that the estimates have high efficiency relative to exact binomial and resulting inference procedures usually adequate for practical application.

2- Total counts of identically independent distributed (i.i.d.) binary variables (equivalently, sums of i.i.d. binary variables coded as 1 or 0) follow a binomial distribution, and a model of zero-inflated regression is useful when there are a high proportion of zero counts in the data.

3- Beta distribution has an essential application when dealing with binary outcomes because it assigns positive probability for the values no more than 0 and 1. Additionally, beta distribution provides tractable mathematical calculation for binary outcomes. For the values α> 1 and β > 1, the density of a beta distribution with (α and β) as parameters is unimodal.

4- As an alternative model to beta-binomial in the case of overdispersion, the negative binomial regression can be used since the reality of this model is to analyze the over-dispersed data.

4.Conclusions:

The binomial distribution is known for its wide application to medical and biological data. It is always known that there is a change in the pathological response, which leads to instability of probability ratios, and hence the problem of overdispersion. Although there are alternative ways to solve this problem here, beta-binomial is considered as one of the alternative models. Also, to the extent of the importance and necessity of the issue, different types of cancer in the world, including breast cancer, it appeared that our society was not deprived of this deadly epidemic as it appeared in the estimation of event rates. We recommend. In the clinical study, an equal sample size can be considered revising "Maximum Likelihood" estimation in multiple different samples.

Conflict of Interest

None.

References

1. Anderson, D. A.: Some models for overdispersed binomial data, *Aust. J. Stat.*, 30, 125–148, (1988).
2. Kleinman, J. C.: Proportions with Extraneous Variance: Single and Independent Samples, *J. Am. Stat. Assoc.*, 68, 46–54, (1973).
3. Ennis, D. M. and Bi, J.: The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests, *J. Sens. Stud.*, 13, 389–412, (1998).
4. Lee, J. C. and Sabavala, D. J.: Bayesian Estimation and Prediction for the Beta-Binomial Model, *J. Bus. Econ. Stat.*, 5, 357, (1987).
5. Lee, J. and Lio, Y. L.: A note on bayesian estimation and prediction for the beta-binomial model, *J. Stat. Comput. Simul.*, 63, 73–91, (1999).
6. Chuang-Stein, C.: An Application of the Beta-Binomial Model to Combine and Monitor Medical Event Rates in Clinical Trials, *Drug Inf. J.*, 27, 515–523, (1993).
7. Young-Xu, Y. and Chan, K. A.: Pooling overdispersed binomial data to estimate event rate, *BMC Med. Res. Methodol.*, 8, 58, (2008).
8. Griffiths, D. A.: Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease, *Biometrics*, 29, 637, (1973).
9. Guimarães, P.: A simple approach to fit the beta-binomial model, *Stata J.*, 5, 385–394, (2005).
10. Garren, S. T., Smith, R. L., and Piegorsch, W. W.: On a Likelihood-Based Goodness-of-Fit Test of the Beta-Binomial Model, *Biometrics*, 56, 947–949, (2000).
11. Tripathi, R. C., Gupta, R. C., and Gurland, J.: Estimation of parameters in the beta binomial model, *Ann. Inst. Stat. Math.*, 46, 317–331, (1994).
12. Kapourani, C. A.: Beta Binomial for overdispersion, (2008). [Online]. Available: https://rpubs.com/cakapourani/beta-binomial.
13. Bodhisuwan, W. and Saengthong, P.: The Negative Binomial – Weighted Garima Distribution: Model, Properties and Applications, *Pakistan J. Stat. Oper. Res.*, 16, 1–10, (2020).
14. Azimi, S. S., Bahrami Samani, E., and Ganjali, M.: Random Effects Models for Analyzing Mixed Overdispersed Binomial and Normal Longitudinal Responses With Application to Kidney Function Data of Cancer Patients, *Stat. Biopharm. Res.*, 0, 1–18, (2020).
15. Zhang, Y.-Y., Xie, Y.-H., Song, W.-H., and Zhou, M.-Q.: The Bayes rule of the parameter in (0,1) under Zhang's loss function with an application to the beta-binomial model, *Commun. Stat. - Theory Methods*, 49, 1904–1920, (2020).