GOODNESS OF FIT TEST FOR THE PROPORTION OF SUCCESSES IN BINOMIAL TRIALS AND CONFIDENCE INTERVAL VIA COINCIDENCE: A CASE OF RARE EVENTS

Victor Nijimbere
School of Mathematics and Statistics, Carleton University, Ottawa ON, K1S 5B6, Canada
victornijimbere@gmail.com

ABSTRACT

In this paper, we define the coincidence in the context of binomial trials. We consider the null hypothesis $H_0: p_1 = p_2 = \cdots = p_n = \vartheta$ against the alternative hypothesis $H_1: p_i \neq \vartheta$ for some $i, i = 1,2,\cdots,n$, where p_i is the probability of successes (or proportion) in each performed experiment, and ϑ is a real constant. We also consider that ϑ is small ($\vartheta \to 0$) and the number of experiments is large ($\to \infty$), and make use of Poisson limit Theorem to establish a statistical test to examine these hypotheses. We show that if the null hypothesis is not rejected, then most likely the coincidence is expected to occur, and therefore we compute a confidence interval for ϑ in terms of the generalised hypergeometric function (special function) using the variance of the coincidence (or via coincidence). These results are also written in terms of elementary functions using the asymptotic expansion of the hypergeometric function. The obtained results can, for example, be used in information retrieval, health care, natural language processing, quality control in industries, etc.

Keywords: *Coincidence, Poisson limit theorem, Hypothesis test, Confidence interval, Hypergeometric function, asymptotic evaluation*

1. INTRODUCTION

A discrete random variable X having a binomial distribution with parameters k and p is denoted as $X \sim bin(k,p)$, and its probability mass function (p.m.f) is given by

$$P(X = s) = \binom{k}{s} p^s (1-p)^{k-s}, s = 0,1,2,\cdots, \tag{1}$$

where k is the number of experiments performed, s is the number of successes, p (or proportion) is the probability of successes in any given performance, and

$$\binom{k}{s} = \frac{k!}{(k-s)!s!} \tag{2}$$

represents all possible combinations observed in the outcomes. It can be readily shown that the mean of X is kp while the variance is $kp(1-p)$. This can be found in any book on the basics of probability and statistics [2,9].

The binomial distribution has many applications in science, social science, health care, and engineering [4,5,6,9,10]. Information retrieval, natural language processing (Dunning [4]) which are techniques frequently used to make interactions between humans and computers efficient, and quality controls are typical examples among other applications of the later distribution [5].

To obtain meaningful results whenever the binomial distribution is applied, suitable inferential and statistical approaches must be carefully used. The main question is: ''Is the estimate (estimated value) for the parameter p acceptable to justify the use of the binomial distribution model in (1) and at what extent?". To adequately answer this question, one should first perform goodness of fit tests, and then construct confidence intervals.

The mathematical and statistical analysis developed in this paper can be applied in science, social science, healthcare and engineering. In the present study, we will focus on text analysis which is importantly used in information retrieval and natural language processing [4,10], in order to simplify the description of our methods.

Among different statistical tests that may be conducted in text analysis is the Likelihood Ratio Test which is based on maximizing the likelihood function [2]. A good description of this test that consists of comparing two population proportions p_1 and p_2 in text analysis can be found in Dunning [4]. On the other hand, Wallis [10] constructed Wald confidence interval, Wilson's score confidence interval and Clopper-Pearson interval for the parameter p relevant to text analysis, and performed goodness of fit and contingency tests to evaluate these intervals. Dunning [4] suggested that the use of Poisson distribution may provide some benefits on one hand when p is small, and Wallis [10], in the conclusion of his investigations on another hand, suggested that for skewed parameter p and for large $(p \to 0, k \to \infty)$, Poisson distribution would achieve better results.

Before we proceed to the aims of this work, we should first give a short description of Poisson distribution as it is an important tool that we are going to use. A discrete random variable X is Poisson distributed with parameter λ if its probability mass function (p.m.f) is [2,9]

$$P(X = s) = \frac{e^{-\lambda}\lambda^s}{s!}, s = 0,1,2,\cdots. \tag{3}$$

It is denoted as $X \sim Poi(\lambda)$, and it has a very interesting property that both the mean and variance of the random variable X are equal and are given by λ.

In this paper, we consider a more general situation with n independent studies in which experiments are performed with probabilities of successes $p_1, p_2, p_3, \cdots p_n$. And we are interested in establishing a test statistics for evaluating the null hypothesis

$$H_0: p_1 = p_2 = \cdots = p_n = \vartheta, \tag{4}$$

against the alternative hypothesis

$$H_1: p_i \neq \vartheta \text{ for some } i, i = 1,2,\cdots,n, \tag{5}$$

where ϑ is some real number. Thus, the first goal of this work is to establish a goodness of fit test to examine these hypotheses, and the second goal is to compute a confidence interval (CI) for the parameter ϑ if there is no evidence to reject the null hypothesis H_0.

We will consider that the parameter ϑ is small $(\vartheta \to 0)$ and constant, and $k \to \infty$ as suggested in Wallis [10]. In that case, Poisson distribution is a very good approximation for the binomial distribution [9]. This property is known as Poisson limit theorem or the law of rare events [9]. Using this property, we will show shortly (section 2) that the hypotheses in (4) and (5) are equivalent to the hypotheses

$$\tilde{H}_0: \lambda_1 = \lambda_2 = \cdots = \lambda_n = \theta \tag{6}$$

and

$$\tilde{H}_1: \lambda_i \neq \theta \text{ for some } i, i = 1,2,\cdots,n, \tag{7}$$

where $\lambda_i, i = 1,2,\cdots,n$ are parameters of some Poisson distributions and θ is some real constant.

Once this is done, we will then apply known results for Poisson distribution (see Nijimbere [7]) to obtain new results for the binomial distribution. For instance, in the case of Poisson distribution, Nijimbere [7] established a χ^2 goodness of fit test to examine the hypotheses (6) and (7), and constructed a $100(1-\alpha)\%$ confidence interval (CI) for θ using the variance of the coincidence that we will define later (section 3) in the context of binomial trials. In this paper, a new goodness of fit test to examine the hypotheses in (4) and (5) is carried out, and hence a new $100(1-\alpha)\%$ CI for ϑ is obtained using the variance of the coincidence as in Nijimbere [7].

2. APPROXIMATION OF THE BINOMIAL DISTRIBUTION BY POISSON DISTRIBUTION

In this section, we describe the approximation of the binomial distribution by Poisson distribution. And we show that the hypotheses in (4) and (5) and those in (6) and (7) are equivalent. In fact, this is possible when the probability of success is small and the number of trials is large.

Theorem 1. (Poisson limit theorem) *As $p \to 0$ and $k \to \infty$, such that the mean value $kp = \lambda$ is constant, the approximation*

$$\binom{k}{s} p^s (1-p)^{k-s} \approx \frac{e^{-\lambda}\lambda^s}{s!}, s = 0,1,2,\cdots \tag{8}$$

holds.

The proof can be found in [9], we repeat it here.

Proof.

$$\binom{k}{s} p^s (1-p)^{k-s} = \frac{k(k-1)(k-2)\cdots(k-s+1)}{s!}\left(\frac{\lambda}{k}\right)^s \left(1-\frac{\lambda}{k}\right)^{k-s}. \tag{9}$$

And if k is large, $k \to \infty$,

$$\binom{k}{s} p^s (1-p)^{k-s} \approx \frac{k^s}{s!}\left(\frac{\lambda}{k}\right)^s \left(1-\frac{\lambda}{k}\right)^{k-s} = \frac{\lambda^s}{s!}\left(1-\frac{\lambda}{k}\right)^{k-s} = \frac{\lambda^s}{s!}\frac{\left(1-\frac{\lambda}{k}\right)^k}{\left(1-\frac{\lambda}{k}\right)^s} = \frac{e^{-\lambda}\lambda^s}{s!} \tag{10}$$

which is Poisson distribution probability mass function since

$$\left(1-\frac{\lambda}{k}\right)^k \to e^{-\lambda} \quad \text{and} \quad \left(1-\frac{\lambda}{k}\right)^{-s} \to 1 \text{ as } \quad k \to \infty. \tag{11}$$

Moreover, we observe that if $p \to 0$ and $k \to \infty$, then the variance of X is approximated by $kp(1-p) \simeq kp$ which is the mean of X, and that a distribution whose mean and variance are equal is Poisson distribution (see section 1 or [2]). Hence, under this assumption, the random variable X is Poisson distributed with mean $\lambda = kp$, $X \sim Poi(\lambda = kp)$.

Let us now consider that $X_i \sim bin(k_i, p_i), i = 1,\cdots,n$. If $k_i = k$, for all $i = 1,\cdots,n$, then we have from (6) and (7) that

$$\tilde{H}_0: kp_1 = kp_2 = \cdots = kp_n = \theta \tag{12}$$

and

$$\tilde{H}_1: kp_i \neq \theta \text{ for some } i, i = 1,2,\cdots,n. \tag{13}$$

This gives

$$\tilde{H}_0: p_1 = p_2 = \cdots = p_n = \frac{\theta}{k} = \vartheta \tag{14}$$

and

$$\tilde{H}_1: p_i \neq \frac{\theta}{k} = \vartheta \text{ for some } i, i = 1,2,\cdots,n, \tag{15}$$

which are exactly (4) and (5).

3. COINCIDENCE, PROBABILITY AND MOMENTS

In this section, the probability of the coincidence and the moments associated with the coincidence are derived in terms of the hypergeometric function following Nijimbere [7]. But before these derivations, we should first define the coincidence in the context of binomial trials and the generalized hypergeometric function.

Definition 1. *Let $X_i \sim bin(k_i, p_i), i = 1,2,\cdots,n$ be independent and identically distributed (iid). Then, a coincidence will occur if, after counting, the number of successes is the same in all n cases. Thus, the coincidence is given by*

$$C = \bigcup_{s=0}^{\infty} \{X_1 = X_2 = \cdots = X_n = s\}. \tag{16}$$

Definition 2. *The generalized hypergeometric function is a special function given by the power series [1,8]*

$${}_pF_q(a_1, a_2, \cdots, a_p; b_1, b_2, \cdots, b_q; x) = \sum_{s=0}^{\infty} \frac{(a_1)_s (a_2)_s \cdots (a_p)_s}{(b_1)_s (b_2)_s \cdots (b_q)_s} \frac{x^s}{s!}, \tag{17}$$

where $a_1, a_2, \cdots a_p$ and $b_1, b_2, \cdots b_q$ are arbitrary constants, $(\gamma)_s = \frac{\Gamma(\gamma+s)}{\Gamma(\gamma)}$ for any complex γ, with $(\gamma)_0 = 1$, and Γ is the gamma function.

Theorem 2. *Let k_i=k for all $i, i = 1,2,\cdots,n$ (see Definition 1). Then under H_0, and as $\vartheta \to 0$ and $k \to \infty$, the probability of the coincidence C in (12) is approximated by*

$$P(C) \approx e^{-nk\vartheta}\, {}_1F_n(1;1,1,\cdots,1;(k\vartheta)^n). \tag{18}$$

Proof. The joint p.m.f of $X_1 = s_1, X_2 = s_2, \cdots, X_n = s_n$ is the multinomial p.m.f

$$\begin{aligned} &P(X_1 = s_1, X_2 = s_2, \cdots, X_n = s_n) = P(X_1 = s_1)P(X_2 = s_2)\cdots P(X_n = s_n) \\ &= \binom{k_1}{s_1} p_1^{s_1}(1-p_1)^{k_1-s_1} \binom{k_2}{s_2} p_2^{s_2}(1-p_2)^{k_2-s_2} \cdots \binom{k_n}{s_n} p_n^{s_n}(1-p_n)^{k_n-s_n} = \prod_{i=1}^{n} \binom{k_i}{s_i} p_i^{s_i}(1-p_i)^{k_i-s_i}. \end{aligned} \tag{19}$$

If $k_i = k$ for all i,

$$P(X_1 = s_1, X_2 = s_2, \cdots, X_n = s_n) = \prod_{i=1}^{n} \binom{k}{s_i} p_i^{s_i}(1-p_i)^{k-s_i}. \tag{20}$$

Then,

$$P(C) = P(X_1 = X_2 = \cdots = X_n = s) = \sum_{s=0}^{k} \binom{k}{s}^n \prod_{i=1}^{n} p_i^{s}(1-p_i)^{k-s}. \tag{21}$$

Using (4) and (14) and apply Theorem 1 yields

$$P(C) = \lim_{k\to\infty} \sum_{s=0}^{k} \binom{k}{s}^n (\vartheta^n)^s [(1-\vartheta)^n]^{k-s} \approx e^{-n\theta} \sum_{s=0}^{\infty} \frac{(\theta^n)^s}{(s!)^n}, \tag{22}$$

where $\theta = k\vartheta$. Now using Theorem 1 in Nijimbere [7] gives

$$P(C) = \lim_{k\to\infty} \sum_{s=0}^{k} \binom{k}{s}^n (\vartheta^n)^s [(1-\vartheta)^n]^{k-s} \approx e^{-n\theta} \sum_{s=0}^{\infty} \frac{(\theta^n)^s}{(s!)^n} = e^{-n\theta}\, {}_1F_n(1;1,1,\cdots,1;\theta^n). \tag{23}$$

Hence substituting back $\theta = k\vartheta$ gives

$$P(C) \approx e^{-nk\vartheta}\, {}_1F_n(1;1,1,\cdots,1;(k\vartheta)^n). \tag{24}$$

This ends the proof.

Theorem 3. *Under H_0, and as $\vartheta \to 0$ and $k \to \infty$, the γ^{th}moment μ_γ associated with the coincidence C is approximately given by*

$$\mu_\gamma \approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,\cdots,1,2,\cdots,2;(k\vartheta)^n), \tag{25}$$

where $\gamma = 1,2,3,\cdots$. And the variance is approximated by

$$\sigma^2 \approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n) - e^{-2nk\vartheta}(k\vartheta)^{2n}\left[{}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n)\right]^2. \tag{26}$$

Proof. *Under H_0, and as $\vartheta \to 0$ and $k \to \infty$, we have*

$$\mu_\gamma = \lim_{k\to\infty}\sum_{s=0}^{k} s^\gamma \binom{k}{s}^n (\vartheta^n)^s[(1-\vartheta)^n]^{k-s} \approx e^{-n\theta}\sum_{s=0}^{\infty} s^\gamma \frac{(\theta^n)^s}{(s!)^n}, \tag{27}$$

where, as before, $\theta = k\vartheta$. Following Lemma 1 in [7] yields

$$\mu_\gamma = \lim_{k\to\infty}\sum_{s=0}^{k} s^\gamma \binom{k}{s}^n (\vartheta^n)^s[(1-\vartheta)^n]^{k-s} \approx e^{-n\theta}\sum_{s=0}^{\infty} s^\gamma \frac{(\theta^n)^s}{(s!)^n} = e^{-n\theta}\theta^n {}_1F_n(1;1,\cdots,1,2,\cdots,2;\theta^n). \tag{28}$$

Hence, substituting back $\theta = k\vartheta$ gives

$$\mu_\gamma \approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,\cdots,1,2,\cdots,2;(k\vartheta)^n), \tag{29}$$

which is exactly (25).

Now, we can apply Theorem 3 in [7] to obtain

$$\mu_1 = \mu_{\gamma=1} \approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n) \tag{30}$$

and

$$\mu_2 = \mu_{\gamma=2} \approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n). \tag{31}$$

Hence,

$$\begin{aligned}\sigma^2 &= \mu_2 - (\mu_1)^2 = \mu_{\gamma=2} - (\mu_{\gamma=1})^2 \\ &\approx e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n) - e^{-2nk\vartheta}(k\vartheta)^{2n}\left[{}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n)\right]^2.\end{aligned} \tag{32}$$

This completes the proof.

4. GOODNESS OF FIT TEST AND CONFIDENCE INTERVAL (CI) FOR ϑ VIA COINCIDENCE

In this section, a goodness of fit test to examine the hypotheses (4) and (5) is obtained. And if there is no evidence to reject the null hypothesis H_0 in (4), then a $100(1-\alpha)\%$ CI for ϑ is constructed using the Central Limit Theorem (CLT).

Theorem 4. *If k_i=k for all $i, i = 1,2,\cdots,n$ (see Definition 1), $\vartheta \to 0$ and $k \to \infty$, and there is no evidence to reject the null hypothesis H_0, then*

$$P(H_0 := true) = P(C), \tag{33}$$

where $P(C)$ is given by (14).

Proof. If $k_i = k$ for all $i, i = 1,2,\cdots,n$, we have

$$\begin{aligned} &P(H_0 := true) = P(p_1 = p_2 = \cdots = p_n) \\ &= P\left(\frac{X_1}{k_1} = \frac{X_2}{k_2} = \cdots = \frac{X_n}{k_n}\right) = P\left(\frac{X_1}{k} = \frac{X_2}{k} = \cdots = \frac{X_n}{k}\right) = P(C), \end{aligned} \tag{34}$$

where $P(C)$ is given by (1.4) if $\vartheta \to 0$ *and* $k \to \infty$ (Theorem 1). This ends the proof.

One may understand Theorem 4 this way. If the null hypothesis is not rejected, we shall expect more and more coincidences to take place as we perform more and more experiments.

Moreover, if the null hypothesis H_0 in (4) is not rejected, then most likely the coincidence is expected to happen (Theorem 4), and hence the variance of $X_1, X_2, \cdots, X_n$ is that of the coincidence σ^2 given in Theorem 3. And if $k\vartheta \geq 10$, then, by the Central Limit Theorem [3], we have

$$X_1, X_2, \cdots, X_n \sim N(k\vartheta, \sigma^2), \tag{35}$$

where σ^2 is given in Theorem 3.

Let now $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_n$ be the sample estimates for $p_1, p_2, \cdots, p_n$. And let

$$\hat{\vartheta} = \frac{\hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_n}{n}. \tag{36}$$

Therefore,

$$\frac{\vartheta - \hat{\vartheta}}{\left(\frac{\sigma}{\sqrt{k}}\right)} \sim N(0,1), \tag{37}$$

or

$$\frac{\vartheta - \hat{\vartheta}}{\sqrt{\frac{e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n) - e^{-2nk\vartheta}(k\vartheta)^{2n}\left[{}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n)\right]^2}{k}}} \sim N(0,1). \tag{38}$$

Having in mind that $Z \sim N(0,1)$, we can now conduct a test as following. The null hypothesis H_0 will be rejected if

$$\frac{\vartheta - \hat{\vartheta}}{\sqrt{\frac{e^{-nk\vartheta}(k\vartheta)^n {}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n) - e^{-2nk\vartheta}(k\vartheta)^{2n}\left[{}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n)\right]^2}{k}}} < Z_{\frac{\alpha}{2}}, \tag{39}$$

and (or)

$$\frac{\vartheta-\hat{\vartheta}}{\sqrt{\frac{e^{-nk\vartheta}(k\vartheta)^n{}_1F_n(1;1,1,2,\cdots,2;(k\vartheta)^n)-e^{-2nk\vartheta}(k\vartheta)^{2n}\left[{}_1F_n(1;1,2,\cdots,2;(k\vartheta)^n)\right]^2}{k}}} > Z_{1-\frac{\alpha}{2}}. \quad (40)$$

In the case, the null hypothesis is not rejected, a $100(1-\alpha)\%$ CI for ϑ can be computed. It is, in fact, given by

$$\hat{\vartheta}-Z_{\frac{\alpha}{2}}\hat{\sigma}<\vartheta<\hat{\vartheta}+Z_{\frac{\alpha}{2}}\hat{\sigma}, \quad (41)$$

where

$$\hat{\sigma}\approx\sqrt{e^{-nk\hat{\vartheta}}\left(k\hat{\vartheta}\right)^n{}_1F_n\left(1;1,1,2,\cdots,2;\left(k\hat{\vartheta}\right)^n\right)-e^{-2nk\hat{\vartheta}}\left(k\hat{\vartheta}\right)^{2n}\left[{}_1F_n\left(1;1,2,\cdots,2;\left(k\hat{\vartheta}\right)^n\right)\right]^2}, \quad (42)$$

and $\hat{\vartheta}$ is given by (31).
In the case $n=2$, $\hat{\sigma}$ can be expressed in terms of modified Bessel functions of the first kind of order 0 and 1 [7]. Setting $\theta=k\vartheta$ in equation (A.67) in Corollary 1 in [7], we have

$$\sigma^2\approx e^{-2k\vartheta}(k\vartheta)^2I_0(2k\vartheta)-e^{-4k\vartheta}(k\vartheta)^2\left[I_1(2k\vartheta)\right]^2, \quad (43)$$

where I_0 and I_1 are the modified Bessel functions of the orders 0 and 1 respectively [1]. This gives

$$\hat{\sigma}\approx\sqrt{e^{-2k\hat{\vartheta}}\left(k\hat{\vartheta}\right)^2I_0\left(2k\hat{\vartheta}\right)-e^{-4k\hat{\vartheta}}\left(k\hat{\vartheta}\right)^2\left[I_1\left(2k\hat{\vartheta}\right)\right]^2}, \quad (44)$$

where $\hat{\vartheta}=(\hat{p}_1+\hat{p}_2)/2$.

5. FURTHER ASYMPTOTIC EXPENSIONS OF THE CI FOR ϑ

As mentioned in Definition 2, hypergeometric and Bessel functions are special functions. They have very interesting mathematical properties which can be used to simplify the results in section 3. For instance, if $k\hat{\vartheta}\gg 1\ (k\hat{\vartheta}\geq 10)$, one can evaluate $\hat{\sigma}$ in terms of elementary functions rather than spectial functions [1,8]. This is called asymptotic evaluation. The asymptotic expressions for σ^2 were derived in Nijimbere [7].

In the case $n>2$, the asymptotic expressions for σ^2 is given by equation (5.39) in Theorem 6 in [7]. Substitute $\theta=k\vartheta$ in (5.39) in [7], and then substitute the resulting expression for σ^2 in (35), we obtain simpler expressions for (39) and (40). Hence, we should reject the null hypothesis H_0 if

$$\frac{\vartheta-\hat{\vartheta}}{\sqrt{\frac{2(2\pi)^{1/2-n/2}n^{-1/2}\left(k\vartheta\right)^{5/2-n/2}-4n^{-1}(2\pi)^{1-n}\left(k\vartheta\right)^{3-n}}{k}}}<Z_{\frac{\alpha}{2}},$$

(45)

and (or)

$$\frac{\vartheta-\hat{\vartheta}}{\sqrt{\frac{2(2\pi)^{1/2-n/2}n^{-1/2}\left(k\vartheta\right)^{5/2-n/2}-4n^{-1}(2\pi)^{1-n}\left(k\vartheta\right)^{3-n}}{k}}}>Z_{1-\frac{\alpha}{2}}, \quad (46)$$

where as before $\hat{\vartheta}$ is given by (36).

If , by this test, we can not reject the null hypothesis, we can then construct a $100(1-\alpha)\%$ CI for ϑ using (41), and where

$$\hat{\sigma} \sim \sqrt{2(2\pi)^{1/2-n/2} n^{-1/2} \left(k\hat{\vartheta}\right)^{5/2-n/2} - 4n^{-1}(2\pi)^{1-n}\left(k\hat{\vartheta}\right)^{3-n}}\,. \tag{47}$$

In the case $n = 2$, the asymptotic expressions for σ^2 is given by equation (A.73) in Theorem 8 in [7]. Substitute $\theta = k\vartheta$ in (A.73) in [7], and then substitute the resulting expression for σ^2 in (35), we obtain simpler expressions for (39) and (40). Hence, the null hypothesis H_0 shoud be rejected if

$$\frac{\vartheta - \hat{\vartheta}}{\sqrt{\frac{1}{k}}\sqrt{\frac{(k\vartheta)^{3/2}}{2(\pi)^{1/2}} - \frac{k\vartheta}{4\pi}}} < Z_{\frac{\alpha}{2}}\,, \tag{48}$$

and (or)

$$\frac{\vartheta - \hat{\vartheta}}{\sqrt{\frac{1}{k}}\sqrt{\frac{(k\vartheta)^{3/2}}{2(\pi)^{1/2}} - \frac{k\vartheta}{4\pi}}} > Z_{1-\frac{\alpha}{2}}, \tag{49}$$

where as before where $\hat{\vartheta} = (\hat{p}_1 + \hat{p}_2)/2$.
If there is no evidence to reject it, one may construct a $100(1-\alpha)\%$ CI for ϑ using (41), and where

$$\hat{\sigma} \sim \sqrt{\frac{\left(k\hat{\vartheta}\right)^{3/2}}{2(\pi)^{1/2}} - \frac{k\hat{\vartheta}}{4\pi}}. \tag{50}$$

6. DISCUSSION AND CONCLUSION

We defined the coincidence (Definition 1) and computed its probability of occurrence in terms of the hypergeometric function using Poisson limit theorem (Theorem 2). We have used Poisson limit theorem to express the variance and moments of the coincidence in terms of hypergeometric function (Theorem 3).
We further showed that the probability that H_0 is true equals that of the coincidence (Theorem 4). This can be understood this way. If the null hypothesis is true (cannot be rejected), then more coincidences will occur as we keep performing the experiment many times. Therefore, the variance of $X_i, i = 1,2,\cdots,n$ is given by the variance of the coincidence C. In this case, one may use the CLT to establish a statistical test as described in section 4.
Hypergeometric and Bessel functions are special functions having interesting mathematical properties that need some understating of rigorous mathematics. For simplification purpose, asymptotic expansions of these non-elementary functions were used to express the variance of the coincidence C in terms of elementary functions (section 5).
The outcomes of this work can, for instance, be applied to achieve better results in health care, computational linguistics, quality controls, computer science and so on.

7. REFERENCES

[1] M. Abramowitz, I.A. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables". Nat. Bur. Stands (1964).
[2] M.H. DeGroot, M.J. Schervish, "Probability and Statistics", 3rd Ed., Addison-Wesley (2002).
[3] G. Casella, R.L. Berger , "Statistical Inference", 2nd Ed., Duxbury (2001).
[4] T. Dunning, Accurate Method for the Statistics of Surprise and Coincidence, *Computational Linguistics* **19**, Issue 1, 61-74 (1993).
[5] Engineering Statistics Handbooks. www.itl.nist.gov/div898/handbook/pmc/section3/pmc331.htm
[6] M.O. Finkelstein, B. Levin, "Statistics for Lawyers", 3rd Ed., Springer (2015)
[7] V. Nijimbere, Coincidences, Goodness of Fit Test and Confidence Interval for Poisson Distribution Parameter via Coincidence, *American Journal of Applied Mathematics and Statistics* **4**, Issue 6, 185-193 (2016).
[8] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/15
[9] A. Papoulis, S.U. Pillai, "Probability, Random Variables and Stochastic Processes". 4th Ed., McGraw Hill (2002).
[10] S. Wallis, Binomial Confidence Interval and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods, *Journal of Quantitative Linguistics* **20**, Issue 3, 178-208 (2013).