Chapter 2

Review of Literature

2.1 Review of the Related Literature

In this chapter we review some relevant works on the problem of confidence interval construction for the ratio of two binomial proportions. Various authors have come up with different solutions to this problem. Now we provide a brief discussion of the literature pertaining to this subject in order to compare our results with those already known.

The first easily computed methods of confidence estimation for the ratio of binomial proportions have been suggested by Noether (1957) and Guttman (1958). A review of these early methods may be found in Sheps (1959). Methods of confidence estimation of the ratio of proportions as a diagnostic test that can detect a disease are used in McNeil et al. (1975).

Next, methods based on the corresponding significance test have been developed. For example, Thomas and Gart (1977) suggested to apply the method based on fixed marginals in the two-by-two tables for confidence interval construction. Santher et al. (1980) developed and generalizes this method and suggested three related exact methods for finding such intervals.

Katz et al. (1978) suggests three methods of lower confidence limit for the ratio of binomial proportions, and the limits are defined as solutions of some equations. Numerical comparison of the methods shows that the method in which the logarithmic transformation is applied to the ratio of estimates of probabilities is preferential. Some modifications of these methods, that take their origin in Fieller’s method, are discussed in Bailey (1987).

Santher and Snell (1980) derives exact intervals for the risk ratio from Cornfield's (1956) confidence interval for the odds ratio.

Koopman (1984), as well as Miettinen and Nurminen (1985) proposed methods based on the asymptotic likelihood for hypothesis $\theta = \theta_0$ testing with the alternative of $\theta \neq \theta_0$. In Koopman(1984) this method was compared with the one recommended by Katz et al. (1978).

All these results discovered until the end of the ’80 ’s were summarized in Gart and Nam (1988). In this paper they provide a comprehensive survey of various approximation methods of confidence limit constructions for the ratio of probabilities based on the properties of goodness of fit with Pearson's chi-square test, invariance, universality of an application for all observations and computational simplicity. Also, methods for asymptotic of coverage probabilities improving by taking into account the asymptotic asymmetry of statistics are suggested (see also Gart and Nam (1990)). The results obtained are extended for the case of estimating the common ratio in a series of two-by-two tables, which was considered before in Gart (1985). Extensive numerical illustrations are provided, which allow to compare accuracy properties of the methods of interval estimation of probability ratios. Instead of iterative algorithms for calculating the approximate confidence intervals that have been provided by Koopman (1984), Gart and Nam (1988), Nam (1995) give the analytical solutions for upper and lower confidence limits in a closed form.

For interval estimator construction, Bedrick (1987) uses the special power divergence family of statistics. Intervals based on inverting the Pearson, likelihood-ratio, and Freeman-Tukey statistics are included in this family. Asymptotic efficiency, coverage probability, and mean interval length are investigated. Comparisons of methods are provided by numerical examples.

The bootstrap method of a confidence interval construction for the ratio of binomial proportions is suggested in Kinsella (1987)

Coe and Tamhane (1993) provided a method for small sample confidence interval construction for the difference of probabilities, based on an extension of Sterne's method known for constructing small sample confidence intervals for a single success probability. Modifications of the algorithm for ratio probabilities are also indicated.

Nam and Blackwelder (2002) developed a superior alternative to the Wald’s interval and gave corresponding sample size formulas. Bonett and Price (2006) proposed alternatives to the Nam-Blackwelder confidence interval based on combining two Wilson score intervals. Two sample size formulas are derived to approximate the sample size required to achieve an interval estimate with desired confidence level and length.

Extensive numerical illustrations for comparison of exact and asymptotic methods for the ratio of binomial proportions confidence interval construction are presented in the thesis by Mukhopadhyay (2003).

A common epidemiological problem is concerned with estimating the ratio, θ, of the proportion of patients who respond in two treatment groups, or two groups of a cohort study. The response may, in fact, be the manifestation of a side effect, and θ is often known as the risk ratio. In this paper we are concerned with the construction of large-sample confidence intervals for θ.

In this paper we present a new method, allied to that recommended by Katz et al. (1987), and so providing easily computed limits.

As it is mentioned in Bailey's (1987) paper, for the problem of obtaining confidence limits to the risk ratio, or the ratio of two binomial probabilities, he proposes a method based on a power of the observed ratio, the power being chosen to minimize the skewness of the pivotal random variable. The method is simple to use and more stable than that of the previously known. A continuity correction is suggested if a conservative interval is desired.

We note once again that although many statisticians have studied in more details the ratio of two binomial proportions (see all the papers mentioned above), to the best of our knowledge no one has studied the problem of constructing confidence intervals for the ratio of proportions for inverse sampling. Our research considers both of direct binomial sampling and also inverse sampling. Our goal is to compare the performance of all five new methods with the previously known.

In the thesis we consider two different schemes of Binomial sampling. Direct Binomial sampling: Sample size is fixed before the experiment. Inverse binomial sampling: The number of successes in the sample is fixed, while sample size is random. The sampling stops after the last m^{th} success is achieved.

2.2 Confidence limits with using direct and inverse binomial sampling methods

Sufficiently simple methods of asymptotic confidence limits construction for the ratio of probabilities $\theta = p_1/p_2$ exist in the case when the stopping moment for observations from the Bernoulli sequence with success probability p_1 are priory fixed $(\upsilon_1 = n)$. That is, the observations are done in the framework of direct binomial sampling, while observations from the sequence with success probability p_2 are done as inverse binomial sampling. That is, the stopping time υ is defined by the number of the observation that results in achieving $m(\geq 1)$ successes.

The likelihood function of the random samples $\left(X^{(n)}, Y^{(\nu)}\right)$ depends on the components of these samples only through the values of complete sufficient statistics $\left(\sum_1^n X_k, \upsilon\right)$. The distribution of the statistic $T = \sum_1^n X_k$ follows the Binomial law $B(n, p_1)$, and the distribution of υ follows the Negative Binomial distribution (sometimes also called Pascal law) $Nb(m, p_2)$. It is well known that statistic $\overline{X}_n = T/n$ has the mean value $\mu_1 = p_1$, variance $\sigma_1^2 = p_1(1-p_1)/n$, and is asymptotically $(n \to \infty)$ normal with parameters (μ_1, σ_1^2). Statistic $\overline{Y}_m = \upsilon/m$ has the mean value $\mu_2 = 1/p_2$, variance $\sigma_2^2 = (1-p_2)/mp_2^2$, and is asymptotically $(m \to \infty)$ normal with parameters (μ_2, σ_2^2) (Gut (1995), Apendix).

Hence (see Lehmann (1998), Chapter 2, Section 1), $\hat{\theta}_{n,m} = \overline{X}_n \overline{Y}_m$ is an unbiased estimation of probabilities ratio θ such that uniformly by all values of p_1, p_2 it minimizes any risk function with convex loss function and it is asymptotically $(n, m \to \infty)$ normal with mean $\mu = \theta$ and variance

$$\sigma^2 = \frac{p_1(1-p_1)}{p_2^2 n} + \frac{p_1^2(1-p_2)}{p_2^2 m} = \theta\left[\frac{p_2^{-1} - \theta}{n} + \frac{\theta - p_1}{m}\right] \text{-----------------(1)}$$

The last statement immediately follows from the following easy to prove lemma.

Lemma 1. Let X_n be asymptotically $(n \to \infty)$ normal $(\mu_1, \sigma_1^2/n)$ and Y_m be asymptotically $(m \to \infty)$ normal $(\mu_2, \sigma_2^2/m,)$, then $X_n \cdot Y_m$ is asymptotically

$(n,m \to \infty)$ normal with parameters $\mu = \mu_1 \mu_2$ and $\sigma^2 = \mu_2^2 \sigma_1^2 / n + \mu_1^2 \sigma_2^2 / m$.

Proof. Introduce a normalized random variable

$$Z_{n,m} = \frac{X_n - \mu_1}{\sigma_1}\sqrt{n} \cdot \frac{Y_m - \mu_2}{\sigma_2}\sqrt{m},$$

which under simultaneous limits n and m to infinity has a nondegenerate distribution. Then

$$X_n \cdot Y_m = Z_{n,m} \cdot \frac{\sigma_1 \sigma_2}{\sqrt{nm}} + X_n \mu_2 + Y_m \mu_1 - \mu_1 \mu_2.$$

Hence, by Slutsky's theorem, the asymptotic distribution of $X_n \cdot Y_m$ coincides with the asymptotic distribution of $X_n \mu_2 + Y_m \mu_1 - \mu_1 \mu_2$.

Below three confidence intervals for the direct-inverse case are obtained. In all formulas the standard procedure of obtaning normal approximation is used: Point estimation plus/minus $Z_{\alpha/2}$ times standard error. In formula(3) the asymptotic variance from Lemma 1 is used. Contrary to this, in formula(4) the true value of the variance is substituted. Formula(6) is based on completely different idea than formula(3) and (4). For (6) the number of successes is fixed and equals to the number of successes in the first sample. The results obtained allow us to the state the following theorem.

Theorem 1. If $n, m \to \infty$, then an asymptotic $(1-\alpha)$-confidence region (interval) for the parametric function θ as defined by the following inequality

$$\left|\theta - \hat{\theta}_{n,m}\right| \le Z_{\alpha/2}\sqrt{\theta\left(\frac{\overline{Y}_m - \theta}{n} + \frac{\theta - \overline{X}_n}{m}\right)}. \text{ -----------------------------(2)}$$

The interval with bounds

$$\hat{\theta}_{n,m} \pm Z_{\alpha/2}\sqrt{\overline{X}_n \overline{Y}_m \left(\frac{\overline{Y}_m (1 - \overline{X}_n)}{n} + \frac{\overline{X}_n (\overline{Y}_m - 1)}{m}\right)} \text{ ---------------------(3)}$$

is an asymptotically $(1-\alpha)$-confidence interval for θ, where $Z_{\alpha/2}$ is the quantile of standard normal distribution.

Proof. The statements follow from the asymptotic normality of the estimate $\hat{\theta}_{n,m}$. If in the right hand side of formula (1) for the asymptotic variance of the estimate we change p_1 and p_2^{-1} on their consistent estimates $\overline{X}_n$ and $\overline{Y}_m$ respectively, then we obtain the asymptotically confidence region (2). If additionally in (1) we change θ on its estimate $\hat{\theta}_{n,m}$, then we obtain the confidence interval (3).

Note that the left and right bounds of intervals (2) and (3) are the asymptotically lower and upper $(1-\alpha/2)$-confidence bounds for the parametric function θ.

While if the true, not asymptotic variance of $\hat{\theta}_{n,m}$ is used, then all characteristics of the confidence interval direct-inverse sampling, confidence interval becomes better. Hence, the following ("upgraded") variant of Theorem 1 is suggested.

Theorem 2. If $n,m \to \infty$, then an confidence interval for the parametric function θ with bounds

$$\hat{\theta}_{n,m} \pm Z_{\alpha/2}\hat{\theta}_{n,m}\sqrt{\overline{X}_n\overline{Y}_m\left(\frac{(\overline{Y}_m-1)(1-\overline{X}_n)}{nm}+\frac{\overline{Y}_m(1-\overline{X}_n)}{n}+\frac{\overline{X}_n(\overline{Y}_m-1)}{m}\right)}, \text{----}(4)$$

is an asymptotically $(1-\alpha)$-confidence interval for θ, where $Z_{\alpha/2}$ is the quantile of standard normal distribution.

The important part of the suggested realization of the estimate of θ is the choice of the number m. The (random) sample size for the second sample depends on this number. If a statistician could obtain at least the same size of sample n which she had in the first sample, then the following sampling plan for the second stage of the statistical experiment can be suggested. Repeat observations until the same number of successes as in the first experiment, that is, set $m = T$. Of course, we consider only the case when the value of T is greater than zero. Then for the estimate of $1/p_2$ it is natural to consider the statistics $\overline{Y}_T = \upsilon/T$, where the conditional distribution of υ is the Negative Binomial distribution $Nb(T, p_2)$ and the unconditional distribution is obtained by taking the expectation of this distribution by the truncated at zero Binomial distribution T. The estimate of the parameter θ is $\hat{\theta}_n = \overline{X}_n\overline{Y}_m = \frac{T}{n}.\frac{\upsilon}{m} = \frac{\upsilon}{n}$ because $T = m$.

Lemma 2. If $n \to \infty$, then the estimate $\hat{\theta}_n$ Is asymptotically normal with the mean $\mu = \theta$ and variance

$$\sigma^2 = \frac{\theta}{n}\left(\frac{2}{p_2} - \theta - 1\right).$$

Proof. The characteristic function of the Negative Binomial distribution $Nb(m, p_2)$ (the distribution of υ given $T = m$) is $\varphi_m(t) = \lambda^m(t)$, where

$$\lambda(t) = \frac{p_2 \mathrm{e}^{\mathrm{i}t}}{1 - (1 - p_2)\mathrm{e}^{\mathrm{i}t}}.$$

Under the assumption that T has truncated at zero Binomial distribution, the characteristic function of the unconditional distribution of ν takes the form

$$\varphi(t) = \frac{1}{1-(1-p_1)^n} \cdot \sum_{i=1}^{n} \genfrac{}{}{0pt}{}{n}{i} \; [p_1 \lambda(t)]^i (1-p_1)^{n-i}$$

$$= \frac{[p_1 \lambda(t) + (1-p_1)]^n - (1-p_1)^n}{1-(1-p_1)^n}.$$

The statement of the lemma follows now from the Taylor expansion of the function $\varphi(t)$. The lemma immediately implies the following result.

Theorem 3. If $n \to \infty$, the asymptotic $(1-\alpha)$-confidence interval for the parametric function θ is defined by the inequality

$$\left|\theta - \hat{\theta}_n\right| \le Z_{\alpha/2}\sqrt{\frac{\theta}{n}\left(2\overline{Y}_T - \theta - 1\right)}\text{------------------------(5)}$$

The interval bounded by the points

$$\hat{\theta}_n \pm Z_{\alpha/2}\sqrt{\frac{\hat{\theta}_n}{n}\left(2\overline{Y}_T - \hat{\theta}_n - 1\right)}\text{---------------------------(6)}$$

is the asymptotically $(1-\alpha)$-confidence interval for θ, where $Z_{\alpha/2}$ is the quantile of standard normal distribution.

Of course, the left and right ends of the intervals (3), (4), and (6) provide the asymptotically upper and lower $(1-\alpha/2)$ confidence boundaries for the parametric function θ

2.3 Confidence limits with using only direct binomial sampling method

Consider now a standard situation when a statistician has in his hands only the numbers of success

$$n\overline{X}_n = \sum_1^n X_i, \qquad m\overline{Y}_m = \sum_1^m X_i$$

for two binomial experiments $B(n,p_1)$ and $B(m,p_2)$ with priory fixed sample sizes n and m. Initially for such type of data, asymptotic confidence intervals were constructed on the basis of the statistic of sample means ratio $\overline{X}_n/\overline{Y}_m$. That is, for $1/p_2$, a biased estimation was explored. Moreover, a problem with its irregular behaviour under the absence of successes in trials $B(m,p_2)$ appear. As it has been mentioned in introduction, in this case there is no unbiased estimation of the parametric function $1/p_2$. But it is possible to construct an estimation of this function that has an exponentially small value of a bias for $m \to \infty$.

Let $X_1,...,X_n$ be a sample in Bernoulli scheme with success probability p, and $T = \sum_1^n X_i$. For a construction of an estimate $\hat{\theta}_n$ of the parametric function $\theta = 1/p$, we apply the statistic υ, which equals to the number of the last trial with $X_\upsilon = 1$. Then, by the analogy with the inverse binomial sampling, it is natural to suggest the statistic $\hat{\theta}_n = \upsilon/T$ as the estimate of θ. But the value of υ in our case is unknown, so it is better to use the projection $\theta_n^* = \theta_n^*(T) = E\{\hat{\theta}_n|T\}$ of this statistic on the sufficient statistic T. As it is known, (see Lehmann (1998), Chapter 2, Section 1), a projection does not cause an increase of the risk if the loss function is convex.

Lemma 2. The projected estimator has the following representation $\theta_n^* = (n+1)/(T+1)$ and its mean value is

$$E[\theta_n^*(T)] = \frac{1}{p}\left(1-(1-p)^{n+1}\right).$$

Proof. The joint distribution of statistics ν and T is defined by the probabilities

$$P(\upsilon = k, T = t) = \begin{cases} 0, & if \quad k = 0, t \geq 1, \\ (1-p)^n, & if \quad k = 0, t = 0, \\ \binom{k-1}{t-1} p^t (1-p)^{n-t}, & if \quad t = 1,\ldots, n, k = t, \ldots, n. \end{cases}$$

The marginal distribution of statistic T is

$$P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}, \quad t = 0, 1, \ldots, n,$$

then the conditional distribution

$$P(\upsilon = k | T = t) = \begin{cases} 0, & if \quad k = 0, t \geq 1, \\ 1, & if \quad k = 0, t = 0, \\ \binom{k-1}{t-1} / \binom{n}{t}, & if \quad t = 1, \ldots, n, k = t, \ldots, n. \end{cases}$$

All further calculations for mean values are trivial, if we use the well known combinatorial formula

$$\sum_{k=1}^{N} \binom{n+k}{k} = \binom{n+N+1}{n+1}.$$

It follows from the lemma proved above that for an estimate of the parametric function $\theta = p_1/p_2$, it is appropriate to take the statistic

$$\hat{\theta}_{n,m} = \frac{\overline{X}_n (m+1)}{m\overline{Y}_m + 1},$$

with mean value

$$E[\hat{\theta}_{n,m}] = \theta\left(1 - (1-p_2)^{m+1}\right).$$

The next theorem provides two kinds of asymptotic confidence intervals for θ.

Theorem 4. If $n, m \to \infty$, then an asymptotic $(1-\alpha)$-confident region (interval) for the parametric function θ is defined by the inequality

$$\left|\theta - \hat{\theta}_{n,m}\right| \leq Z_{\alpha/2} \sqrt{\theta\left(\frac{1-\overline{X}_n}{n\overline{Y}_m} + \theta \frac{1-\overline{Y}_m}{m\overline{Y}_m}\right)}. \text{---------------------------}(7)$$

The interval with bounds

$$\hat{\theta}_{n,m} \pm Z_{\alpha/2}\sqrt{\hat{\theta}_{n,m}\left(\frac{1-\overline{X}_n}{n\overline{Y}_m} + \hat{\theta}_{n,m}\frac{(1-\overline{Y}_m)}{m\overline{Y}_m}\right)} \text{-------------------------(8)}$$

is an asymptotically $(1-\alpha)$-confident interval for θ , where $Z_{\alpha/2}$ is the quantile of standard normal distribution.

Proof. By the analogy with the proof of Theorem 1, the current proof follows from the following asymptotic representation (the standard technique of asymptotic normality parameters' calculations for a ratio of two asymptotically normal estimates is used):

$$\hat{\theta}_{n,m} = \frac{\overline{X}_n}{p_2} - \frac{1}{p_2}\left[XY\sqrt{\frac{p_1 p_2(1-p_1)(1-p_2)}{nm}} + p_1(\overline{Y}_m - p_2) + O_{p_2}\left(\frac{1}{m^2}\right)\right],$$

where

$$X = \frac{\overline{X}_n - p_1}{p_1(1-p_1)}\sqrt{n}, \quad Y = \frac{\overline{Y}_m - p_2}{p_2(1-p_2)}\sqrt{m}\,.$$

Therefore, the confidence interval constructed above is asymptotically equivalent to the interval based on the statistic $\overline{X}_n/\overline{Y}_m$, but the problem that appears when the denominator of the estimate is zero with a positive probability is completely solved, and the estimate with smaller bias is explored. An interested reader may compare with a solution of this problem in the paper Cho (2007); see the beginning of Section 2.

2.4 Confidence limits with using only inverse binomial sampling method

For the case when both samples are obtained in the schemes $Nb(m_i, p_i),\ i = 1,2$ of the inverse binomial sampling, there exists an unbiased estimate of θ with the uniformly minimal risk for any loss function. Really, for the parametric function $1/p_2$ we have the unbiased estimate υ_2/m_2, and for p_1 under the scheme of inverse sampling for $m_1 \geq 2$ there also exists the unbiased estimate (see Guttman, I. (1958)) $\hat{p}_1 = (m_1 - 1)/(\upsilon_1 - 1)$. Therefore, the optimal unbiased estimate of $\theta = p_1/p_2$ is

$$\hat{\theta}_{n,m} = \frac{\upsilon_2(m_1 - 1)}{(\upsilon_1 - 1)m_2}.$$

We have that $E[\upsilon_i/m_i] = 1/p_i$, $Var[\upsilon_i/m_i] = (1 - p_i)/m_i\, p_i^2$, $i = 1,2$, and by the same method of asymptotic analysis for a ratio of two asymptotically normal estimates that we explored in the previous section, we obtain the following theorem.

Theorem 5. If $m_i \to \infty, i = 1,2$, then the interval bounded by the points

$$\hat{\theta}_{n,m} \pm Z_{\alpha/2}\sqrt{\hat{\theta}_{n,m}\left(\frac{\hat{p}_1(1 - \hat{p}_2)}{m_2} + \hat{\theta}_{n,m}\frac{1 - \hat{p}_1}{m_1}\right)} \text{-------------------------------}(9)$$

where

$\hat{p}_i = (m_i - 1)/(\upsilon_i - 1), i = 1,2,$ is an asymptotically $(1 - \alpha)$ confidence interval for θ , and $Z_{\alpha/2}$ is the quantile of standard normal distribution.