Chapter 1

Introduction

1.1 Statement of the Problems and Importance of the study

The ratio of two binomial proportions arises in prospective studies, biological experiments, or comparison of manufacturing processes for quality control in industry. Statistical procedures for the ratio of binomial proportions (often called the relative risk) is also quite common in clinical trials and epidemiological study, or more generally in the pharmaceutical setting. In epidemiological problems, such as cohort studies in two groups, the risk ratio or odds ratio is related to vaccine efficiency and attributable risk.

Next, in various public health applications, people are interested in knowing whether certain pollutants in the air might increase the chances of a disease by two fold. In clinical trials one of the main goals is to test whether the new drug performs better than an existing drug (or placebo) for curing a certain disease. The problem can be set up as (say), whether the new drug has a cure rate of 1.5 times the existing drug (or placebo). Or if we are evaluating the disease incidence rate (more appropriate in vaccine trials), then people would be interested in testing whether the new treatment reduces the chances of occurrence of the disease by (say) 50%.

Note that in all these applications the problem may be set up in the following way. We have two groups, say the treatment group, and the placebo group. Our responses are dichotomous (such as, a success or a failure). If we denote by p_1 and p_2 the true proportions of successes in each of the groups, then we are interested in estimation of the relative risk $\theta = p_1/p_2$.

For the various applications mentioned above, confidence intervals for the relative risk are typically provided along with the estimate. This gives a good idea about the bounds within which the true value of the relative risk parameter θ will lie.

Since the responses in each group are binary, one can use the binomial distribution to model the data. Constructing confidence intervals for the discrete

distribution setting has been an old problem of interest, and to the best of our knowledge, there has not been any right or wrong method of doing this. One simple minded approach would be to ignore the discreteness completely, i.e., assume asymptotic normality and construct asymptotic intervals. The problem with this approach is that there is no guarantee that the coverage probability of the interval will be at least above a nominal level. The coverage probability gives the chance of an interval containing the true value of the parameter when the experiment is repeated for a large number of times. Typically one wants the coverage probability to be at least greater than or equal to a certain nominal coverage (such as 95%). The problem with the asymptotic intervals are that they may be very liberal and the coverage probability may fall far short of the nominal coverage.

The confidence intervals that are computed based on the actual distribution (instead of any asymptotic normality assumption) are typically called "Exact Confidence Intervals". These intervals may be constructed in such a manner that the coverage can be guaranteed to be at least $100(1-\alpha)\%$. The usual method of constructing such intervals is to invert an equal tailed test.

In many applications, such as in clinical trials or vaccine trials, or more generally in the pharmaceutical setting this guaranteeing of nominal coverage is a desirable property. However the main problem with the exact intervals is that they tend to be very conservative, especially for smaller sample sizes. By more conservative, we mean that the coverage probability is strictly greater than the nominal coverage and the lengths of the interval are very wide. Because of that we would still like to consider not exact, but asymptotic confidence intervals.

Up to our knowledge, this statistical problem has been solved only for schemes of sampling with a fixed number of observations. That is, for direct binomial sampling schemes. For example, in Koopman (1984), a method was proposed for obtaining approximate confidence intervals for the ratio of two binomial proportions based on two independent binonmially distributed random variables. For the problem of obtaining confidence limits to the risk ratio, or the ratio of two binomial probabilities, Katz et al. (1978) proposed a method based on the logarithm of the observed ratio. Bailey (1987) proposed a method based on a power of the observed ratio, the power being chosen to minimize the skewness of the pivotal random variable.

Although many statisticians have studied in more details the ratio of two binomial proportions, to the best of our knowledge no one has studied the problem of constructing confidence intervals for the ratio of proportions for inverse sampling; while our research considers not only the case of direct binomial sampling, but also the inverse. The inverse binomial sampling is the basis of the sequential Wald's analysis. It provides at least double gain (twice less) in the required sample size for making a decision with fixed boundaries on accuracy and reliability of statistical inference. It is also frequently used for a quality control of a production, and for a decision of acceptance of a new medical treatment method in comparison with an old one, when the observations (patients or animals) are arriving in a sequence.

A complexity of the problem stated can be explained by two reasons. Generally speaking, the difficulties arise because none of the known techniques of confidence interval construction work in this situation.

It is well known that the construction of a confidence intervals is closely connected with hypothesis testing. Roughly speaking we can say that a confidence interval is nothing more than the "acceptance region" for the corresponding hypothesis testing procedure. But there is no good (uniformly most powerful) test for hypothesis testing of $\theta = \theta_0$ in the case of an arbitrary hypothetical value θ_0. As it is known, the uniformly most powerful unbiased test exists for the values of the cross-product ratio $\rho = p_1(1-p_2)/p_2(1-p_1)$, but this is not what we need. Hence, it seems to be impossible to use the standard method of confidence interval construction based on an acceptance region of the corresponding hypothesis test.

Another, but not less important difficulty, is that it is not possible to apply the pivotal quantity method. This is the most common and well established technique of confidence interval construction and corresponding hypothesis testing. Unfortunaltely, for the problem of estimation of the ratio of two binomial proportions, there is no pivot function with the desirable accuracy properties. This happens because of the absence of an unbiased estimation for the parametric function $1/p$ for Bernoulli trials with the fixed sample size n, that is, for the direct sampling method (see Lehmann (1998), Chapter 2, Section 1). But if the inverse, not direct binomial selection method is used, then such unbiased estimation exists. This is the starting point of our research

for the construction of confidence intervals for a ratio of success probabilities. The main novel features of this thesis is that we would like to compare the five new methods of confidence intervals.

1. The confidence limits using the direct and inverse binomial sampling methods. There are 3 cases in this method as follows
    1.1 when the asymptotic variance is used.
    1.2 when the true value of the variance is used.
    1.3 when the number of successes is fixed as in the first experiment.
2. The confidence limits using only the direct binomial sampling method.
3. The confidence limits using only the inverse binomial sampling method.

Moreover, we would like to compare the new method for only direct binomial sampling with the previously known.

1.2 Research Objectives

In this thesis, we will compare the coverage probability median, mean and standard deviation of intervals lengths. That is, we compare five new confidence intervals suggested in my thesis according to the sampling scheme (both direct, both inverse, one direct and another inverse). We also compare the performance of the new direct-direct case confidence interval suggested in this thesis with the previously known from the liturature.

1.3 Research Hypothesis

The coverage probabilities of five new confidence intervals are close to nominal coverage for most value of binomial proportions.

The new confidence intervals have sufficiently small mean and median length.

The coverage probabilities of the new confidence intervals for direct binomial sampling are closer to nominal level than previously known.

1.4 Research Scope

The scope of the research includes the following practical results:

1. This section requires the statistical modeling of the true confidence level and nominal. Consider only at the 95% confidence level

2. For each of five types of new confidence intervals proposed in this thesis (both for direct sampling, both for inverse sampling, and combined) run simulations independently and report mean coverage probability, median, mean and standard deviation of the interval length.

3. The value for probabilities of success are .1,.3,.5,.7,.9 (for the simplicity we consider only the case $p_2 \geq p_1$), and three levels of sample size n,m =30,50,100 ,but the case of direct and inverse method when we fix the number of success as in the first experiment, the sample size n=30,50,70,100

4. The values of m_1 and m_2 for the inverse sampling should be calculated by formulae $m_1 = n_1 * p_1$ and $m_2 = n_2 * p_2$.

5. Simulations are repeated 10,000 times for calculating the coverage probabilitity, median, mean and standard deviation of interval lenghts by using the program R version 2.6.1.

1.5 Criteria to Compare

The comparison will be determined by comparing the coverage probabilities, median and mean of the confidence interval lengths.

1.6 Research Advantages

1. This study will present new methods of confidence interval construction for the ratio of two binomial proportions for different sampling schemes.

2. The confidence intervals may be useful for statistical analysis in prospective studies, such as biological experiments, comparison of manufacturing processes for quality control in industry, and etc.

1.7 Benefits of the Research

The benefits of the thesis are as follows:

1. To show the efficiency of new constructions for confidence intervals in comparison with the previously known.

2. To show that five new constructions for confidence intervals are suitable for various types of sampling schemes.