Biostatistics Department Technical Report

BST2009-003

Investigation of the Distribution of the Score Statistic for a Simple Hypothesis in Pool Screening

Charles R. Katholi, PhD
Inmaculada Aban, PhD

Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
ckatholi@uab.edu
caban@uab.edu

Investigation of the Distribution of the Score Statistic For a Simple Hypothesis in Pool Screening

Charles Katholi and Inmaculada Aban

0. Abstract

The Score statistic is one of several statistics based on the use of the likelihood function and its asymptotic properties are key to its applications. The large sample properties of the statistic are well known for the case of independently identically distributed random variables from a continuous distribution. In this report we explore that application of the statistic for testing a simple hypothesis in the case of independent but not identically distributed sampling from a family of discrete distributions. The usual theorems on which asymptotic tests are based are proved and the cumulants of the score statistic examined as a way of assessing the "speed" of convergence to the asymptotic Standard Normal distribution. Simulation methods are used to investigate the true distribution of the score statistic under the null hypothesis and recommendations are made concerning using simulation results to choose critical values for tests. In addition, the Cornish Fisher method is used to obtain critical values for small samples and these are compared to the asymptotic critical values and to the values found by simulation. Proposals are made for the application of the score statistic in practice for the case of pool screening or group testing.

1. Introduction

This technical report is aimed at considering the asymptotic properties of the Score Statistic for testing the simple hypothesis,

$$H_0 : p = p_0 \quad \text{vs} \quad H_A : p \neq p_0$$

in the pool screening approach to estimating the probability of a rare event, *p*. Pool screening (a.k.a. group testing) is often used when the prevalence of infection in an experimental unit is small enough that testing individual units is prohibitively expensive in time and money. From the perspective of infectious diseases, vectors are transmitters of disease-causing organisms that carry pathogens from one host to another. Examples of this are mosquitoes for malaria and black flies for Onchocerciasis. In many disease elimination efforts vector transmitted diseases, success brings with it a greatly reduced population of the vector species that are infected. For example, the infection rates that one wishes to detect might be as low as 5 in 10,000 or less. Thus pool screening based on the identification of segments of DNA from the disease by using polymerase chain reaction (PCR) methods to amplify amounts found in pools of the vector species is an efficient way to test for positivity of the pool and has become a standard way of estimating the prevalence.

To review the notation and state the underlying probability model, let p be the probability that an experimental unit is positive for infection and the value to be estimated and tested against a particular value. Suppose that n independents units are collected and combined in a group or pool, and the pool is then tested in a way that indicates a positive or negative result for the presence of infection. The probability that the pool is not infected is $(1-p)^n$ and hence the probability that it is infected is $1-(1-p)^n$. Denote the result of the test by the Bernoulli random variable X, $X \in \{0,1\}$ .Thus, in pool testing a sample of *m* pools is gathered. The $j-th$ pool is assumed to be of size n_j and the result of testing the $j-th$ pool is denoted by X_j which is a Bernoulli random variables, X_j, $j=1,\cdots,m$ with probability mass function,

$$f_{X_j}(x \mid p, n_j) = \left[1-(1-p)^{n_j}\right]^{X_j}\left[(1-p)^{n_j}\right]^{1-X_j}, \quad X_j \in \{0,1\}, \quad 1 \le n_j \le N_{\max}$$

The X_j, $j=1,\cdots,m$, are independent and members of the same family of distributions. However, that value n_j , the size of the pool tested, can be different for each pool. It is also assumed that there is some upper limit on the size of the pool that can be tested and this upper limit is denoted by $N_{\max}$.

The problems of point estimation and confidence intervals for this situation have been extensively studied.(Barker 2000, Chang and Reeves 1962, Hepworth 1996, Katholi, et. al, 1995, 2006). More recently, Tebbs and McCann (2007) considered Large Sample hypothesis tests for stratified group testing data. They considered the Likelihood Ratio and Wald tests.

In this report we consider the Score Statistic and note that in this case a modest amount of analytic results can be obtained. In particular, we shall show that for small values of the parameter, p_0, very large numbers of pools are required in order for the usual asymptotic results to apply. Following Field and Ronchetti (1990) we evaluate the use of the Edgeworth expansion and the associated Cornish-Fisher expansion for finding p-values and critical values for the test statistic when samples are small. In addition, we examine the use of simulations to estimate the critical values and p-values.

2. The Score Statistic and Its Properties

To this end we note that the log likelihood function in this case is,

$$L(p \mid n_1,\cdots,n_m; X_1,\cdots,X_m) = \sum_{j=1}^{m} X_j \ln\left[1-(1-p)^{n_j}\right] + \sum_{j=1}^{m} n_j (1-X_j)\ln(1-p)$$

So that, after a small amount of algebra, the derivative with respect to *p* of the log likelihood function is,

$$U_m(p \mid X_1,\cdots,X_m;n_1,\cdots,n_m)=\frac{1}{(1-p)}\sum_{j=1}^{m}\left\{\frac{n_jX_j}{[1-(1-p)^{n_j}]}-n_j\right\}$$

also known as the Score Statistic. It follows easily from properties of sums of random variables and the moments of the random variables X_j, $j=1,\cdots,m$ that,

$$E(U_m)=\frac{1}{(1-p)}\sum_{j=1}^{m}\left\{\frac{n_jE(X_j)}{[1-(1-p)^{n_j}]}-n_j\right\}=0$$

since $E(X_j)=[1-(1-p)^{n_j}]$. Similarly, the variance of U_m is,

$$Var(U_m)=E(U_m^2)=$$
$$\frac{1}{(1-p)^2}\left(\sum_{j=1}^{m}E\left\{\frac{n_jX_j}{[1-(1-p)^{n_j}]}-n_j\right\}^2+\sum_{i=1}^{m}\sum_{\substack{k=1\\k\neq i}}^{m}E\left\{\frac{n_iX_i}{[1-(1-p)^{n_i}]}-n_i\right\}\left\{\frac{n_kX_k}{[1-(1-p)^{n_k}]}-n_k\right\}\right)$$

The right hand double sum is equal to zero since each term in the sum is a product of independent random variables, each with expected value equal to zero. Hence,

$$Var(U_m)=\frac{1}{(1-p)^2}\left(\sum_{j=1}^{m}E\left\{\frac{n_j^2X_j^2}{[1-(1-p)^{j}]^2}-2\frac{n_j^2X_j}{[1-(1-p)^{n_j}]}+n_j^2\right\}\right)$$
$$=\frac{1}{(1-p)^2}\left(\sum_{j=1}^{m}\left\{\frac{n_j^2}{[1-(1-p)^{n_j}]}-n_j^2\right\}\right)=\frac{1}{(1-p)^2}\sum_{j=1}^{m}\frac{n_j^2(1-p)^{n_j}}{[1-(1-p)^{n_j}]}$$

which we note is just Fisher's Expected Information.

We shall examine the statistic the standardized score statistic,

$$Z_m=\frac{U_m(p_0 \mid X_1,\cdots,X_m;n_1,\cdots,n_m)}{\sqrt{Var(U_m(p_0))}} \qquad (1)$$

and will show by means of Liapounov's central limit theorem that $Z_m \xrightarrow{D} N(0,1)$ as $m\to\infty$. Although this result is well known for the case of independently and identically distributed random variables, it must be established *ab initio* in this instance since the random variables X_i, although independent, are not identically distributed unless $n_1=n_2=\cdots=n_m$.

Theorem 1: $Z_m \xrightarrow{D} N(0,1)$ as $m\to\infty$

Proof of Theorem 1: Define $Y_k = \frac{1}{(1-p)}\left(\frac{n_k X_k}{[1-(1-p)^{n_k}]} - n_k\right)$ so that from above we know that

$$E(Y_k) = 0 \text{ and } Var(Y_k) = \frac{n_k^2(1-p)^{n_k}}{(1-p)^2[1-(1-p)^k]} = \sigma_k^2 .$$

Let $s_m^2 = \sum_{j=1}^{m} \sigma_j^2$ and define $\nu_{2+\delta}^{(k)} = E\left|Y_k - E(Y_k)\right|^{2+\delta}$ and let $\rho_m = \frac{\sum_{k=1}^{m} \nu_{2+\delta}^{(k)}}{\left[\sqrt{s_m^2}\right]^{2+\delta}}$ where $0 < \delta \le 1$. If we can show that $\lim_{m\to\infty} \rho_m = 0$ then by Liapounov's Central Limit Theorem we will have proved that $Z_m \xrightarrow{D} N(0,1)$ as $m \to \infty$. To begin, we note that

$$E\left|Y_k - E(Y_k)\right|^{2+\delta} = \sum_{X_k \in \{0,1\}} [1-(1-p)^{n_k}]^{X_k}[(1-p)^{n_k}]^{(1-X_k)}\left|\frac{n_k X_k}{(1-p)[1-(1-p)^{n_k}]} - n_k\right|^{2+\delta}$$

$$= \frac{(1-p)^{n_k}}{(1-p)^{2+\delta}}\left|-n_k\right|^{2+\delta} + \frac{[1-(1-p)^{n_k}]}{(1-p)^{2+\delta}}\left|\frac{n_k}{[1-(1-p)^{n_k}]} - n_k\right|^{2+\delta}$$

Some simple algebra leads to the expression,

$$\nu_{2+\delta}^{(k)} = \frac{n_k^{2+\delta}(1-p)^{n_k}}{(1-p)^{2+\delta}}\left[1 + \frac{(1-p)^{n_k(1+\delta)}}{[1-(1-p)^{n_k}]^{1+\delta}}\right]$$

Next we make some general observations which will allow us to bound the $\nu_{2+\delta}^{(k)}$.

Observation 1: For any n_j and n_k such that $n_j \le n_k$ and $0 < p < 1$,

(a). $1 > (1-p)^{n_j} \ge (1-p)^{n_k} > 0$ and $0 < [1-(1-p)^{n_j}] \le [1-(1-p)^{n_k}] < 1$

(b). $\frac{1}{[1-(1-p)^{n_j}]} \ge \frac{1}{[1-(1-p)^{n_k}]}$

Observation 2: Let $n_{(1)} \le n_{(2)} \le \cdots \le n_{(m)}$ be the ordered values of $n_1, \cdots, n_m$ the for all j,

$$\frac{(1-p)^{n_{(1)}}}{[1-(1-p)^{n_{(1)}}]} \ge \frac{(1-p)^{n_j}}{[1-(1-p)^{n_j}]} \ge 0$$

From these two observations we can make the following assertion,

$$\nu_{2+\delta}^{(k)} \le \frac{n_{(m)}^{2+\delta}(1-p)^{n_{(1)}}}{(1-p)^{2+\delta}}\left[1+\left\{\frac{(1-p)^{n_{(1)}}}{[1-(1-p)^{n_{(1)}}]}\right\}^{1+\delta}\right] = B < \infty, \text{for all k}$$

This will be true for *p* in any closed interval inside the open interval (0,1).

Hence we have shown that $\sum_{k=1}^{m}\nu_{2+\delta}^{(k)} \le mB$. Similarly, we can show that

$s_m^2 = \sum_{k=1}^{m}\sigma_k^2 \ge \frac{mn_{(1)}^2(1-p)^{n_{(m)}}}{(1-p)^2[1-(1-p)^{n_{(m)}}]}$ so that the quantity $\rho_m = \frac{\sum_{k=1}^{m}\nu_{2+\delta}^{(k)}}{\left[\sqrt{s_m^2}\right]^{2+\delta}}$ is bounded by,

$$\rho_m \le \frac{1}{m^{\delta/2}}\left[\frac{n_{(m)}^2(1-p)^{n_{(1)}}[1-(1-p)^{n_{(m)}}]^{1+\delta/2}}{[n_{(1)}(1-p)^{n_{(m)}}]^{1+\delta}}\right] \le \frac{1}{m^{\delta/2}}\left[\frac{N^2(1-p)[1-(1-p)^N]^{1+\delta/2}}{[(1-p)^N]^{1+\delta}}\right] < \infty$$

since by assumption there is a number $N = N_{\max} \ni 1 \le n_k \le N_{\max}$, $\forall k$. The quantity in the brackets on the right side of the inequality is independent of *m* and so $\rho_m = \bigcirc\left(m^{-\frac{\delta}{2}}\right)$ as $m \to \infty$. This completes the proof. □

Alternate proof of Theorem 1: It should be noted that an alternate proof is available in this case based on the following theorem which is proved using Liapounov's theorem.

Theorem: Let $X_k, k > 1$ be independent random variables such that $P(a \le X_k \le b) = 1$ for some finite scalars *a* and *b* with $a < b$. Also let $E(X_k) = \mu_k, Var(X_k) = \sigma_k^2$, $T_n = \sum_{k=1}^{n} X_k$, $\xi_n = \sum_{k=1}^{n}\mu_k$ and $s_n^2 = \sum_{k=1}^{n}\sigma_k^2$. Then $Z_n = (T_n - \xi_n)/s_n \xrightarrow{D} N(0,1)$ if and only if $s_n \to \infty$ as $n \to \infty$. Since the random variable $Y_k = \frac{1}{(1-p)}\left(\frac{n_k X_k}{[1-(1-p)^{n_k}]} - n_k\right)$ can take on only two values corresponding to $X_k = 0$ and $X_k = 1$, Y_k can only take on the values $a_k = \frac{-n_k}{(1-p)}$ and $b_k = \frac{n_k(1-p)^{n_k}}{(1-p)[1-(1-p)^{n_k}]}$. These two values still depend on *k* and so we need bounds which are true for all *k*. To this end, we note that the $n_k \in \{1, 2, \cdots, N\}$ and so $a_k \ge \frac{-N}{(1-p)}$, $\forall k$. Examining b_k we observe that it can be written as

$$b_k = \frac{n_k(1-p)^{n_k-1}}{[1-(1-p)][1+(1-p)+\cdots+(1-p)^{n_k-1}]}$$

$$= \frac{n_k}{p[1+(1-p)^{-1}+(1-p)^{-2}+\cdots+(1-p)^{-(n_k-1)}]} \le \frac{1}{p}$$

Since each of the n_k terms in the sum in the denominator is greater than or equal to 1. Thus for all k, $Y_k \in \left[\frac{-N}{(1-p)}, \frac{1}{p}\right] = [a,b]$ and $P(a \le Y_k \le b) = P(Y_k = a_k) + P(Y_k = b_k) = 1$ since $(a_k, b_k) \subseteq (a,b)$, $\forall k$. The same argument used above to obtain bounds shows that $s_n \to \infty$ as $n \to \infty$. Thus, the desired property is demonstrated in either case.□

As with most asymptotic results, it is not clear how large the sample must be for the results to be valid. In this case of the Score Statistic for the pool screening model, the random variable X_k appears in U_m in such a way that taking expectations is very easy and so exploring the distributional properties of the statistic $Z_m = \frac{U_m(p_0 \mid X_1, \cdots, X_m; n_1, \cdots, n_m)}{\sqrt{Var(U_m(p_0))}}$ is feasible. To this end, we shall find the cumulant generating function of Z_m. To begin we note that the moment generating function of X_k is $M_{X_k}(t) = (1-p)^{n_k} + [1-(1-p)^{n_k}]e^t$. Let $\alpha_k = \frac{n_k}{(1-p)[1-(1-p)^{n_k}]}$, $\beta_k = \frac{-n_k}{(1-p)}$ and $\gamma = \sqrt{\text{var}(U_m(p_0))}$ so that $\frac{Y_k}{\gamma} = \frac{\alpha}{\gamma} X_k + \frac{\beta}{\gamma}$. From elementary probability theory we know that the moment generating function of $Y_k \big/ \gamma = e^{\frac{\beta_k}{\gamma} t} M_{X_k}(\frac{\alpha_k}{\gamma} t)$. Finally, the moment generating function for Z_m is equal to

$$M_{Z_m}(t) = e^{t \sum_{j=1}^{m} \beta_j \big/ \gamma} \prod_{k=1}^{m} M_{X_k}\left(\frac{\alpha_k}{\gamma} t\right).$$

Taking the natural log of this expression yields the cumulant generating function. The successive cumulants are found by differentiating with respect to *t* and then evaluating the derivatives at *t = 0*. These computations are feasible through the use of an algebraic programming package like Maple 9.5 for example. This process yields the following formulas for the first 4 cumulants.

$K_1 = 0$

$K_2 = 1$

$$K_3 = \sum_{k=1}^{m}\left(\frac{\frac{n_k^3(1-p)^{n_k}[2(1-p)^{n_k}-1]}{(1-p)^3[1-(1-p)^{n_k}]^2}}{\left\{\sum_{k=1}^{m}\frac{n_k^2(1-p)^{n_k}}{(1-p)^2[1-(1-p)^{n_k}]}\right\}^{3/2}}\right)$$

and

$$K_4 = \sum_{k=1}^{m}\left(\frac{\frac{n_k^4(1-p)^{n_k}[1-6(1-p)^{n_k}+6(1-p)^{2n_k}]}{(1-p)^4[1-(1-p)^{n_k}]^3}}{\left\{\sum_{k=1}^{m}\frac{n_k^2(1-p)^{n_k}}{(1-p)^2[1-(1-p)^{n_k}]}\right\}^{2}}\right)$$

Formulas for cumulants $K_5,\cdots,K_{10}$ are given in the appendix.

We will now show that $K_3 = O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{(1-p)}{p}}\right)$ and $K_4 = O\left(\frac{1}{m}\left[\frac{(1-p)}{p}\right]\right)$ as $m \to \infty$. From this we get a sense of the speed of convergence of the score statistic to a N(0,1) statistic. This also shows us how large the sample must be when p is small in order to be justified in using the asymptotic critical values for the test. To begin we examine the bracketed sum in the denominator of each of these expressions. As we have seen, the quantity

$$\begin{aligned}[1-(1-p)^{n_k}] &= [1-(1-p)][1+(1-p)+\cdots+(1-p)^{n_k-1}] \\ &= p[1+(1-p)+\cdots+(1-p)^{n_k-1}]\end{aligned}$$

If we define the quantity r_k as

$$r_k = \frac{n_k(1-p)^{n_k-1}}{[1+(1-p)+\cdots+(1-p)^{n_k-1}]} < 1 \text{ for } 0 < p < 1$$

then

$$\sum_{k=1}^{m}\frac{n_k^2(1-p)^{n_k}}{(1-p)^2[1-(1-p)^{n_k}]} = \frac{1}{p(1-p)}\sum_{k=1}^{m}n_k r_k \le \frac{m}{p(1-p)}\left[\frac{1}{m}\sum_{k=1}^{m}n_k\right] \le \frac{mN}{p(1-p)}$$

Looking next at the numerator of K_3 we can rewrite it as,

$$\frac{n_k^3(1-p)^{n_k}[2(1-p)^{n_k}-1]}{(1-p)^3[1-(1-p)^{n_k}]^2} = \frac{n_k[n_k(1-p)^{n_k-1}]^2[2(1-p)^{n_k}-1]}{p^2(1-p)^{n_k+1}[1+(1-p)+\cdots+(1-p)^{n_k-1}]^2} = \frac{n_k r_k^2[2(1-p)^{n_k}-1]}{p^2(1-p)^{n_k+1}}$$

So that K_3 is expressed in the more easily understood form,

$$K_3 = \frac{\frac{m}{p^2(1-p)}\left(\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^2[2(1-p)^{n_k}-1]}{(1-p)^{n_k}}\right)}{\left\{\frac{m}{p(1-p)}\left[\frac{1}{m}\sum_{k=1}^{m} n_k r_k\right]\right\}^{3/2}} = \left[\sqrt{\frac{(1-p)}{mp}}\right]\left[\frac{\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^2[2(1-p)^{n_k}-1]}{(1-p)^{n_k}}}{\left\{\frac{1}{m}\sum_{k=1}^{m} n_k r_k\right\}^{3/2}}\right]$$

The terms of the sum inside the numerator of this expression can be bounded as follows: From **Observation 1**, we know that $(1-p)^{-n_k} \le (1-p)^{-N_{\max}}, \forall k.$ It has been previously shown that $0 < r_k < 1$ and it is easily shown that $\left|2(1-p)^{n_k}-1\right| \le 1, 0 \le p \le 1.$ Thus the individual terms in the sum are all less than or equal to $n_k(1-p)^{-N_{\max}} < N_{\max}(1-p)^{-N_{\max}}$ and so the average of these terms is also bounded by $N_{\max}(1-p)^{-N_{\max}}$ which is finite for $0 \le p < 1.$ Since pool screening is only appropriate when p is small (i.e. when the event of interest is rare say less than one in 1000) and so this bound will not be too large. Similarly, the denominator is well bounded and so the large sample behavior depends essentially on the term $\left[\sqrt{\frac{(1-p)}{mp}}\right]$. Considering the numerator sum in K_4 we have

$$\frac{n_k^4(1-p)^{n_k}[1-6(1-p)^{n_k}+6(1-p)^{2n_k}]}{(1-p)^4[1-(1-p)^{n_k}]^3} = \frac{n_k r_k^3[1-6(1-p)^{n_k}+6(1-p)^{2n_k}]}{(1-p)(1-p)^{2n_k}\,p^3}$$

Thus K_4 can be written as,

$$K_4 = \frac{\frac{m}{p^3(1-p)}\left[\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^3[1-6(1-p)^{n_k}+6(1-p)^{2n_k}]}{(1-p)^{2n_k}}\right]}{\left\{\frac{m}{p(1-p)}\left[\frac{1}{m}\sum_{k=1}^{m} n_k r_k\right]\right\}^2}$$

$$= \left[\frac{(1-p)}{mp}\right]\frac{\left[\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^3[1-6(1-p)^{n_k}+6(1-p)^{2n_k}]}{(1-p)^{2n_k}}\right]}{\left[\frac{1}{m}\sum_{k=1}^{m} n_k r_k\right]^2}$$

By the same sort of argument as used for K_3 we can conclude that $K_4 = \bigcirc\left(\frac{1-p}{mp}\right)$ as $m \to \infty$. It is shown in the appendix that in general, $K_k = \bigcirc\left[\left(\frac{1-p}{mp}\right)^{\frac{k}{2}}\right]$. If the statistic is to converge to a $N(0,1)$ random variable, the cumulants $K_r(m)$, $r \geq 3$ must converge to zero as $m \to \infty$. This will not be even close to true when *p* is small unless m is very large. In practice, a very large sample is unlikely and so we must explore alternative methods of evaluating an observed test statistic.

3. Approximating the Distribution and Quantile Points

We have already shown that the score statistic converges in law to N(0,1). Generally in practice, the score statistic is squared and the asymptotic distribution of the result is a Chi Square distribution with one degree of freedom. In the situation being considered, this is not recommended because when a statistic is squared, the skewness information is lost. In this section we will examine the size of the cumulants for various pool sizes and various ranges of the parameter, *p*, with the hope of improving the approximation to the distribution of the score statistic using its cumulants.

3.1 Approximation by Simulation

We will look at how we can use simulations to estimate the quantiles as well as values of the CDF. For continuous distributions, sample estimates of quantile values are generally given in terms of order statistics (Gibbons and Chakrabarti,2003; Sen and Singer,1993). In general, the p-th quantile $Q_X(p)$ is defined to be the smallest value of X at which the CDF is at least equal to p, or

$$Q_X(p) = \inf_X \left(P(X \leq x) \geq p\right) = \inf_X (F_X(x) \geq p)$$

where $F_X(x)$ is the CDF of the random variable X. This definition applies equally well to continuous and discrete distributions. To find the sample quantile, let

$$r = \begin{cases} np \text{ if } np \text{ is an integer} \\ [np+1] \text{ if } np \text{ is not an integer} \end{cases}$$

where $[x]$ denotes the largest integer not exceeding x. Note that different authors give different definitions of the sample p-th quantile. The one given above is that used by Gibbons and Chakraborti (2003). Alternatively, Singer and Sen (1993) offer two formulas;

$$r = [np] + 1 \text{ or } r = [(n+1)p]$$

Once r is found we define the order statistic $X_{(r)}$ to be the p-th sample quantile. It is also possible to give a confidence interval for the p-th quantile in terms of order statistics. To this end we note the following result for continuous distributions:

Theorem 2: A $(1-\alpha)100\%$ confidence interval for the p-th quantile of a continuous random variable is given by $\left(X_{(r)}, X_{(s)}\right)$ where are r and s are integers such that $1 \le r < s \le n$ and

$$P(X_{(r)} < \kappa_p < X_{(s)}) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \ge 1-\alpha$$

where κ_p is the p-th quantile.

This is a well known result (Gibbons and Charaborti, 2003). Many confidence intervals can be calculated from this theorem since there is only one constraint to define the two values r and s. One possibility is to choose r and s so that the interval is the shortest possible. Another alternative is the equal tail area constraint. This leads to two defining equations, namely,

$$\sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \le \frac{\alpha}{2} \quad \text{and} \quad \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \ge 1-\frac{\alpha}{2}$$

Although this result is generally presented only for continuous distributions, Scheffe and Tukey (1945), Frydman and Simon (2007) showed that the resulting interval was correct for discrete distributions in the sense that the interval can be used with confidence at least $1-\alpha$, but that the coverage probabilities in the discrete case may be much larger than the nominal value requested.

In our study of the use of this approach with simulating the quantile points for the score statistic we have used the second of the two ways of choosing *r* suggested by Singer and Sen; that is, $r = \lfloor (n+1)p \rfloor$. These simulations were done at the same time that the calculations previously reported for the Cornish-Fisher expansion approach to selecting quantile points. In all cases, 200,000 "trials" were utilized in the calculations of the sample for estimating the CDF and hence of estimating the quantiles. It is worth noting before presenting the results that because the estimates are based on calculated values of the score statistic, none of the values can lie in the intervals in which the true CDF is flat; they must lie in the intervals as characterized in Lemma 1. For each quantile point estimated, we also checked values on each side of it to see what the frequency of this value was in the simulated data. In addition, 99% confidence intervals for each quantile were calculated as described above, and then the smallest value, $r_1 \le r$ and the largest value $s_1 \ge s$ were found such that $X_{(r_1)} = X_{(r_1+1)} = \cdots = X_{(r)}$ and $X_{(s)} = X_{(s+1)} = \cdots = X_{(s_1)}$. The coverage of the interval $\left[X_{(r_1)}, X_{(s_1)}\right]$ is calculated using the formula in Theorem 1. For a given set of pool sizes, the simulation was repeated five times and the results of these are shown in Table 1.

Table 1: Simulation Results based on 5 replicates assuming $p_0 = 0.0005$, m = 300, and pool sizes between 25 and 50 (with mean pool size 37.28 and variance 52.68).

$\alpha = 0.05$	**Quantile Value**	**Frequency of occurrence**	**99% Confidence Interval**	**Coverage**

Lower	-1.94643	109	-1.94654	-1.52105	0.9960
Upper	1.91858	1	1.91730	2.33819	0.9900
Lower	-1.52105	28	-1.94643	-1.52084	0.9948
Upper	2.33723	1	1.91804	2.33862	0.9900
Lower	-1.52116	11	-1.94643	-1.52084	0.9935
Upper	1.91826	1	1.91719	2.33744	0.9900
Lower	-1.94643	106	-1.94654	-1.52105	0.9982
Upper	2.33723	1	1.91794	2.33883	0.9900
Lower	-1.52126	23	-1.94654	-1.52094	0.9983
Upper	1.92008	1	1.91751	2.33819	0.9900

We first notice that generally the lower critical value is more variable than the upper as indicated by the frequency with which this value occurs in the simulated data. Looking at the details of the simulation we see that in some cases the lower value lies in the interval associated with $T=1$ (lower = -1.94643) while in other cases it lies in the interval associated with $T=2$ (lower = -1.52105). The 99% confidence interval on the other hand always has its lower value in the interval corresponding to $T=1$ and the upper limit in the interval corresponding to $T=2$. The upper quantile value behaves in much the same way and once again the 99% confidence interval has its lower value in the interval corresponding to $T=10$ and the upper value in the interval corresponding to $T=11$. Another set of five simulations were done with a different set of uniformly distributed pool sizes between 25 and 50 and for $p_0=0.0005$. The results were very much the same. The confidence intervals all contained the "flat" region of the CDF although the estimated quantiles were quite noisy.

Based on these simulation results, it would seem that the best recommendation one could make concerning the use of simulated critical values would be to calculate the confidence interval rather than the critical value then make the usual decisions for observed values of the statistic outside the intervals, but to declare values inside the interval as being insufficient to make a decision. We note in passing the p-values are also commonly determined by simulation. These experiments suggest that multiple simulations need to be used to estimate these under H_0 and that some measure of the variability needs to be reported as recently advocated by Koehler, Brown and Haneuse (2009).

3.2 Cornish Fisher Expansion

Each of the cumulant formulas K_k for $k\geq 3$ has a polynomial in powers of $(1-p)^{n_k}$ in its algebraic representation . We will begin by looking at the behavior of these as a function of $x=(1-p)^{n_k}$. We are particularly interested in how large the values of the polynomials can become for $x\in[0,1]$. Since the values of these polynomials contribute to the size of the cumulant depending on the particular value of p, knowledge of their largest possible values helps in bounding the cumulants. The size of the cumulant has a direct bearing on the size of the correction terms in both the Cornish-Fisher and Edgeworth expansions which will be the basis for our calculations. Since both of these

are asymptotic expansions, they are generally not convergent in the usual sense, but a few terms often give good approximations to quantities of interest. Generally the terms get smaller in magnitude for a number of terms and then diverge. The greater the number of decreasing terms the better, since the usual rule is to sum terms until the magnitude starts to increase. The following plot shows that the polynomials have multiple extrema on for $x \in [0,1]$ and that the absolute magnitude of the polynomial at these local maximum and minimum points grows larger as the degrees of the polynomials increase. Also, we note that in practice, these are polynomials in $x = x(p) = (1-p)^{n_k}$ and that p is generally in the range $0 < p < 0.01$.

For $x \in [0,1]$, simple calculations yield the following bounds for values of the polynomials:

Table 2: Bounds for polynomials

Polynomial	Lower Bound	Upper Bound
$f_5(x)$	-0.629	0.629
$f_6(x)$	-0.875	1.000
$f_7(x)$	-1.815	1.815
$f_8(x)$	-4.250	3.900
$f_9(x)$	-10.201	10.201
$f_{10}(x)$	-24.396	31.000

From the table of bounds we see what is clear from the plots in Figure 1a and Figure 1b; that is, that the largest case size of the polynomials grows as the order of the cumulant increases.

Figure 1a: Cumulants K3 through K8

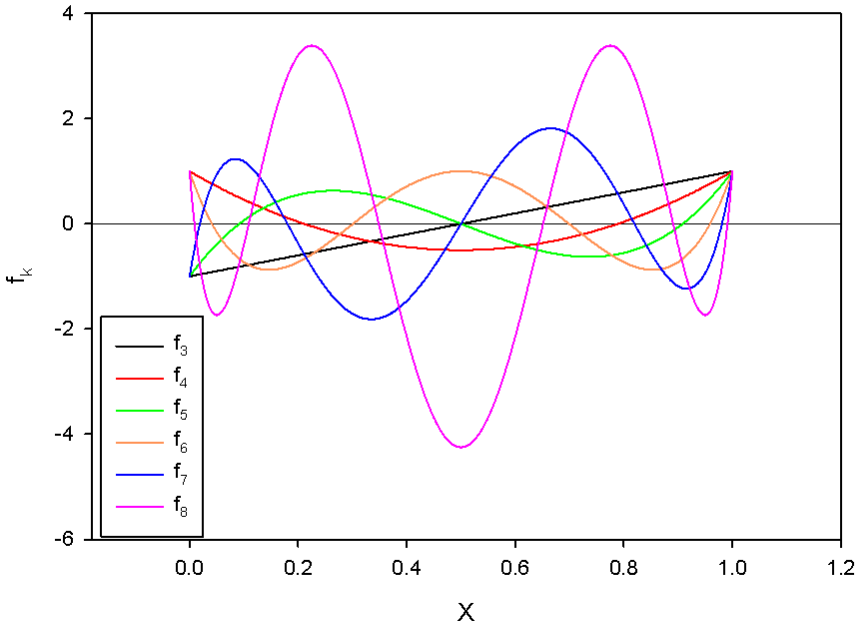
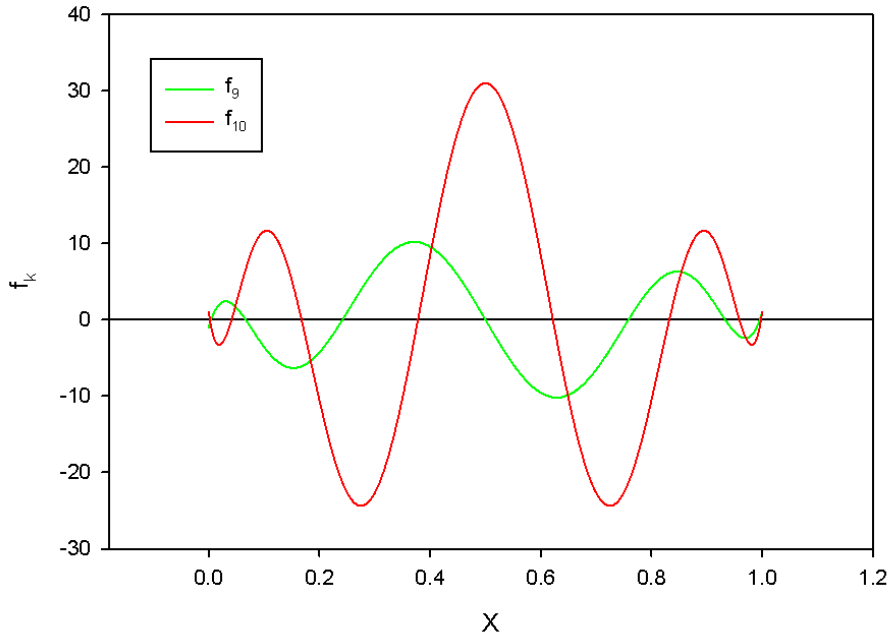


Figure 1b: Cumulants K3 through K8



Next we turn to finding approximate quantiles of the score statistic by means of the Cornish-Fisher expansion. Since we know that the score statistic is asymptotically N(0,1) , the Cornish-Fisher expansion is a possible vehicle for this purpose. The whole idea of this expansion is to find adjustments to the asymptotic normal critical values which incorporate the information about the cumulants of the distribution being

approximated. Recall that the cumulants of the standard normal are all equal to zero after K_2. The Cornish-Fisher expansion is the inversion of the Edgeworth expansion and as such is an approximation to the quantile function of the distribution of the test statistic. It must, however, be used with care since it is an asymptotic expansion and as such does not converge except under special circumstances. On the other hand, given the cumulants of the score statistic, the individual terms of the expansion are easily calculated using the algorithm of Lee and Lee (1992) and the FORTRAN code published by Lee and Lin (1992). These authors suggest calculating the "correction" terms and adding terms as long as they are decreasing in magnitude, then truncating the expansion when the size of the terms starts to grow again. This is the usual strategy for asymptotic expansions. There is one caveat, however, and that is the fact that the score statistic in the case of pool screening is a discrete random variable**.** As we shall see, even when the number of pools is large, the cumulative distribution function (CDF) of the score statistic, especially when p_0 is very small, has a distinctly discrete quality with sections of "flatness" and short intervals of rapid change.

To investigate the performance of the Cornish Fisher approximations, simulations will be used to estimate the distribution of the score statistic for a given value of the parameter, p. The pool sizes also must be specified and in the simulations, these are chosen to follow a discrete uniform distribution on the interval $[n_L, n_H]$, where $n_L \geq 1$ and $n_H \leq N$. Recalling that pool screening is only appropriate when the value of *p* is small, the simulation results will be confined to small values of p (i.e., p in the neighborhood of 0.0005.). Before proceeding, we note a few facts concerning the score statistic in this case.

Lemma 1: Let $T = \sum_{j=1}^{m} X_j$ be the number of positive pools and let the order statistics for the pool sizes be denoted by $n_{(1)} \leq n_{(2)} \leq \cdots \leq n_{(m)}$. The use of the symbol $\leq$ here rather than $<$ is because there are generally tied values. Then, for $T = t$, the largest and smallest values of the score statistic are respectively,

$$Z_m^{\max} = \frac{1}{(1-p_0)\sqrt{Var(U_m(p_0))}} \left\{ \left(\sum_{j=1}^{t} \frac{n_{(m+1-j)}}{[1-(1-p_0)^{n_{(m+1-j)}}]} \right) - \sum_{i=1}^{m} n_{(i)} \right\}$$

and

$$Z_m^{\min} = \frac{1}{(1-p_0)\sqrt{Var(U_m(p_0))}} \left\{ \left(\sum_{j=1}^{t} \frac{n_{(j)}}{[1-(1-p_0)^{n_{(j)}}]} \right) - \sum_{i=1}^{m} n_{(i)} \right\}$$

where $Var(U_m(p_0)) = \frac{1}{(1-p_0)^2} \sum_{j=1}^{m} \frac{n_j^2 (1-p_0)^{n_j}}{[1-(1-p_0)^{n_j}]}$.

Proof: We begin by noting that the quantities in the brackets in these formulas are what determine the size of the statistic. Next we consider the function,

$$h(\varsigma) = \frac{\varsigma}{[1-(1-p_0)^{\varsigma}]}, \varsigma \geq 1$$

We will show that $h(\varsigma)$ is strictly monotone increasing by showing that its derivative is strictly positive. The derivative (after some algebra) is

$$h'(\varsigma) = \frac{1}{[1-(1-p_0)^{\varsigma}]}\left\{1+\frac{\varsigma(1-p_0)^{\varsigma}\ln(1-p_0)}{[1-(1-p_0)^{\varsigma}]}\right\}$$

The quantity $[1-(1-p_0)^{\varsigma}]$ is positive for all $\varsigma \geq 1$ when $0 < p_0 < 1$. Next we write the term in the brackets as,

$$\left\{1+\frac{\varsigma\ln(1-p_0)(1-p_0)(1-p_0)^{\varsigma-1}}{p_0[1+(1-p_0)+(1-p_0)^2+\cdots(1-p_0)^{\varsigma-1}]}\right\} = \left\{1-\frac{\left[(1-p_0)\left|\ln(1-p_0)\right|\right]}{\left[p_0\right]}\frac{\left[\varsigma(1-p_0)^{\varsigma-1}\right]}{\left[\sum_{j=1}^{\varsigma-1}(1-p_0)^j\right]}\right\}$$

and observe that,

$$\frac{\varsigma(1-p_0)^{\varsigma-1}}{[1+(1-p_0)+\cdots(1-p_0)^{\varsigma-1}]} = \frac{\varsigma}{[1+(1-p_0)^{-1}+\cdots(1-p_0)^{-(\varsigma-1)}]} < 1$$

for $0 < p_0 < 1$ since $[1+(1-p_0)^{-1}+\cdots+(1-p_0)^{-(\varsigma-1)}] > \varsigma$. Similarly,

$$\frac{(1-p_0)\ln(1-p_0)}{p_0} = \frac{-[p_0+\frac{1}{2}p_0^2+\frac{1}{3}p_0^3+\cdots](1-p_0)}{p_0} = -(1-p_0)[1+\tfrac{1}{2}p_0+\tfrac{1}{3}p_0^2+\cdots]$$

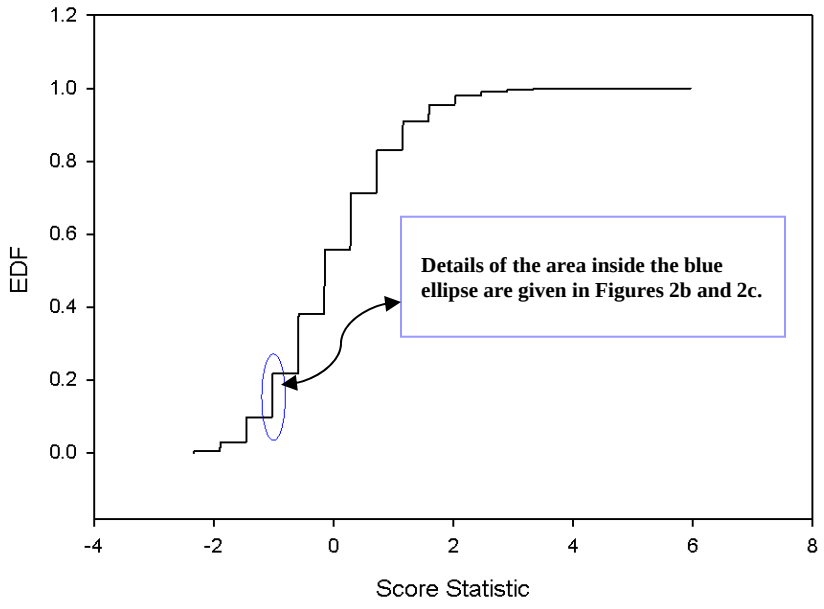
But $[1+\frac{1}{2}p_0+\frac{1}{3}p_0^2+\cdots] < [1+p_0+p_0^2+\cdots] = (1-p_0)^{-1}$ so it follows that

$$C = \frac{\left[(1-p_0)\left|\ln(1-p_0)\right|\right]}{\left[p_0\right]}\frac{\left[\varsigma(1-p_0)^{\varsigma-1}\right]}{\left[\sum_{j=1}^{\varsigma-1}(1-p_0)^j\right]} < 1$$

Thus $1-C > 1-1 = 0$ and so $h(\varsigma)$ is strictly increasing. Returning now to the statistic, Z_m, it follows from Lemma 1, that the largest possible value for the left hand sum occurs when the $n_{(.)}$ correspond to the t largest order statistics and that it is at its smallest for the t smallest order statistics. This completes the proof. □

Lemma 1 shows that the values of the score statistic occur in clusters depending on the observed value of *T*. In particular, for $T = t$, there are possibly as many as $\binom{m}{t}$ values that can occur and these are contained in the interval given in the Lemma. In the special case where all the pool sizes are the same, then each interval collapses into a point since now the distribution of *T* is a binomial with probability of success equal to $[1-(1-p_0)^n]$. Extensive simulations suggest that when the pool sizes are uniformly distributed on [1,N] and the probability of infection is small, the intervals defined by Lemma 1 are disjoint, at least for smaller values of T. As p_0 gets larger or as T gets larger for a fixed p_0, eventually the intervals begin to overlap. As the number of pools gets larger for fixed p_0 the same thing is observed. From these observations we may conclude that it is generally true that the distribution of Z_m has a point probability corresponding to the smallest value (t = 0) and a point probability at its largest value (t=m) although the latter is very unlikely. As the value of *T* increases, the intervals can have larger and larger numbers of values which may occur. It is worth noting in passing that the distribution of *T* can be easily calculated given the values of p_0 and $n_1, n_2, \cdots, n_m$ (Barker 2000). This basic form of the distribution function for Z_m is illustrated for simulated data in Figures 2a and 2b.

Figure 2a: EDF for the Score Statistic Under H0: $p = p_0$
Simulated Data with p = 0.0005
(Pool sizes uniformly in [25,50], Median = 37)



In this figure, each “jump” occurs for a specific observed value of T, the number of positive pools. The transitions here are not, however, a single jump, but are a series of small jumps , one at each of the possible values of the score statistic that can occur when T takes on that specific value. In Figure 2a, the jump corresponding to T = 3 is marked with the blue ellipse. The details of the behavior of the distribution function of the score the EDF of simulated data in the figures 2b and 2c.

Figure 2b:

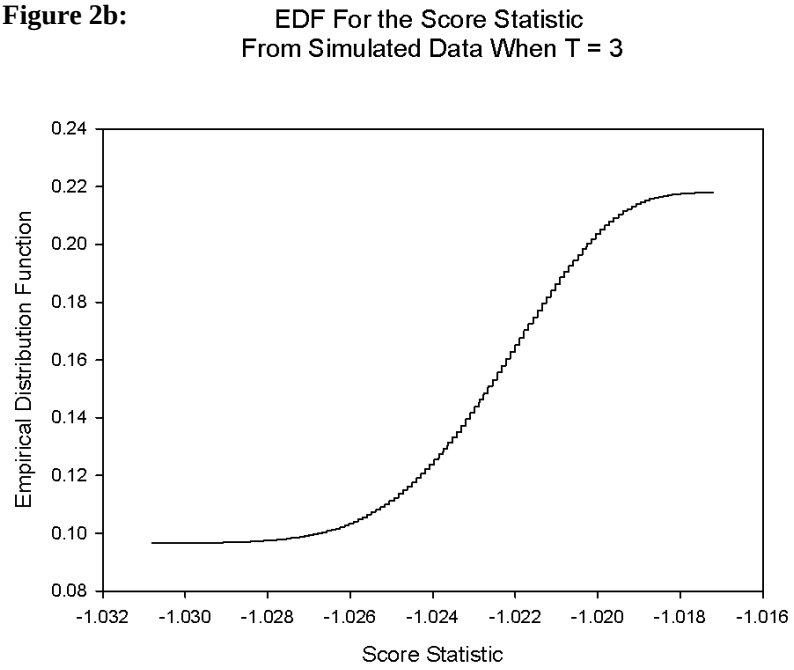
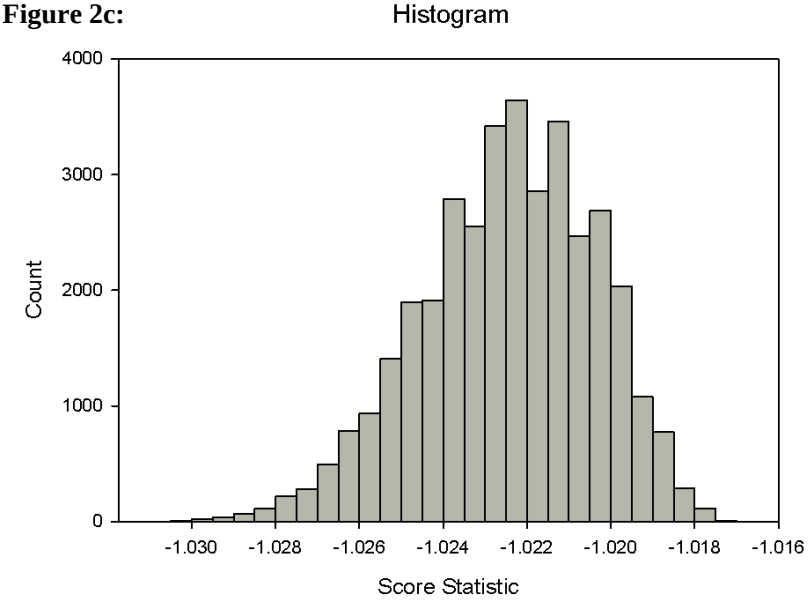


Figure 2c:



For this set of simulated data (i.e., the particular values of the n_j and $p_0 = 0.0005$) the values of the score statistic fall into the interval [-1.03264 , -1.017105] for $T = 3$. As noted previously, for T = t, these are $\binom{m}{t}$ possible ways of assigning which pools are positive out of the m pools. Each of these occurs with a probability which, in principle, can be calculated given p_0 and $n_1, \cdots, n_m$. However, when m is of any size at all, these computations become too extensive for practical work. Thus the size of a jump at any one of the observed simulation points is a reflection of these probabilities through the relative frequency with which the value occurs. The key point to note here is that the distribution of the score statistic is discrete although there can be as many as 2^m values which could occur. Thus, because the values also occur in clusters (related to the value of T) there are also "flat" areas in the distribution function. The choice of $p_0 = 0.0005$ may seem extreme, however, this is exactly the range in which investigators might wish to make a test as part of a disease elimination program where the prevalence is small due to an intervention. Further examples will be given later for larger values of p_0.

With this background information we turn to looking at the use of the Cornish-Fisher expansion as a way of finding critical values when the number of pools is small or moderate. These results will be evaluated by comparison with "true" values found by simulation. The first few terms of the Cornish-Fisher expansion are given, for example, in Abramowitz and Stegun (1964). In the notation of this paper, this becomes,

$$Z_m(p) = x + \left[\frac{1}{6}K_3 He_2(x)\right] + \left[\frac{1}{24}K_4 He_3(x) - \frac{1}{36}K_3^2\left(2He_3(x) + He_1(x)\right)\right] + \cdots$$

where the functions $He_j(x)$ are Hermite polynomials which can be calculated recursively by the formula $He_{k+1}(x) = xHe_k(x) - kHe_{k-1}(x), k \geq 2$ with $He_0(x) = 1$ and $He_1(x) = x$. Note that these polynomials are defined in terms of the probability density function of a standard normal distribution. Thus if we let $Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ then $He_k(x) = (-1)^k Z^{(k)}(x)/Z(x)$. The quantity $Z_m(p)$ is an estimate of the point, $z_m(p)$, at which $P(Z_m \leq z_m(p)) = 1 - p$. The quantity *x* is the point at which the upper tail of the standard normal distribution is equal to p; thus $1 - \int_{-\infty}^{x} Z(\zeta)d\zeta = p$. In performing our simulations, we have used all ten of the cumulants given in the text and the appendix. We have decided where to truncate the expansions by two different methods. First, as described above, terms are added until the next term is greater in absolute value than the current one, and the sum is stopped at this point. Secondly we have used the Levin transformation for summing divergent or slowly convergent series as described in Fessler and Ford (1983) and Smith and Ford (1979,1982). The second approach was utilized to see if it provided, in fact, an automated way of truncating the asymptotic expansion and estimating the error as the Fessler and Ford paper claimed.

The only way to get a sense of the practical usefulness of critical values found by means of the Cornish Fisher expansion is to use simulations to assess the alpha levels which are associated with them. Unfortunately, there are many "moving parts" in the pool screening testing situation and so doing an exhaustive simulation study would take a very large amount of time. Thus, we shall present the results for only a relatively small number of situations that would be sufficient to illustrate the points we want to make. Recall that the things which can change are (1) the number or pools m, (2) the actual pool sizes, n_j (3) the value of p_0 and (4) the alpha level for the test. As we have seen above, the cumulants of the distribution depend only on the first 3 of the items just mentioned and so one set of simulations can be done for each choice of these values to assess the alpha level of a test using the Cornish-Fisher approximate critical values. A second set of simulations can be done for different sets of pool sizes but with a common value for p_0. Finally, sets of simulations can be done for various randomly chosen pool sizes and for different numbers of pools. One objective would be to see at what point the number of pools was sufficiently large that standard asymptotic theory would suffice. In what follows, we will give tables summarizing the results of different simulations. These tables will indicate the number of pools, the range of the pool sizes, the average pool size, the variance of the pool sizes and the value of p_0.

Table 3: Simulated α values assuming $p_0 = 0.0005$, m = 300, and pool sizes between 25 and 50 (with mean pool size 37.28 and variance 52.68).

Simulation Number	Requested α	Observed Lower $\alpha/2$	Observed Upper $\alpha/2$	Observed α
1	0.1	0.02715	0.05426	0.08141
2		0.02726	0.05362	0.08088
3		0.02800	0.05383	0.08183
4		0.02715	0.05376	0.08091
5		0.02759	0.05474	0.08233
Average value		0.02743	0.05404	0.08147
Standard Deviation		0.00037	0.00046	0.00062

Note: The pool sizes were fixed across all five simulations. The upper and lower Cornish-Fisher critical values for the 0.1 level test were respectively 1.75625 and -1.52058.

As has been previously noted, for a given value of p_0 and fixed pool sizes and number of pools, then the possible values of the score statistic lie in disjoint intervals associated with different values of *T*, the number of positive pools. In this case, the value of the upper critical values falls into the interval between *T = 9* and *T = 10*; that is, in intervals where the CDF of the distribution is flat. Hence in this case, since we can calculate the exact probability of getting any value of *T* we can calculate the exact probabilities of being above the critical value. Thus it happens that the probability of a value of the score statistic larger or equal to the upper critical point is equal to $P(T \geq 10) = 0.05367$. The simulated results are, therefore, right on target. The lower critical value in this case fall into the interval associated with *T = 2* and so we cannot use the same approach to judging the adequacy of the simulated values. In evaluating these results it is important to note under what circumstances that a small value of the test

statistic will occur and when a larger value will occur. The answer is a result of the following lemma.

Lemma 2: The function $v(p) = \frac{\varsigma}{[1-(1-p)^{\varsigma}]}$ is a monotone decreasing function of *p*.

Proof: Let $1 > p_a > p_0 > 0$ then it follows that $(1-p_a)^{\varsigma} < (1-p)^{\varsigma}$, $\forall \varsigma \geq 1$. Then it is also true that $1 > 1-(1-p_a)^{\varsigma} > 1-(1-p_0)^{\varsigma} > 0$ and hence that $\frac{\varsigma}{[1-(1-p_a)^{\varsigma}]} < \frac{\varsigma}{[1-(1-p_0)^{\varsigma}]}$, $\forall \geq 1$. Since by hypothesis, $p_a > p_0$ the function is decreasing and monotone. This completes the proof. □

From the lemma, we see that when the true (but unknown) value of *p* is larger than the null hypothesized value, the intervals of statistic values are shifted left; similarly when the true value is less than the hypothesized value the values of the statistic are larger. These observations show that one sided tests can easily be made with the score statistic in this case. Again examining the results above, we note that the actual alpha level when we observe a small value of the test statistic is very conservative while the test for a large value of the test statistic is anti-conservative although not as badly so.

Repeating the same process as described in Table 3 with the alpha level changed from 0.1 to 0.05 yielded Cornish Fisher lower and upper 0.025 level critical values of -1.74851 and 2.14362 respectively. Both of these values fall in the "flat" places in the distribution for the score statistic and so we find that the actual upper and lower alpha levels are 0.02473 and 0.02494; thus the alpha level is 0.04967 which is only slightly conservative and very near the desired level. Note as well, that unlike the case when $\alpha = 0.1$ discussed above, the test is conservative, but not badly so, no matter whether the true value of *p*, is above or below p_0. Finally, when we consider the alpha = 0.01 level test, we find Cornish Fisher critical values of -2.13765 and 2.93342 which again are in "flat" areas of the true distribution function. The associated tail areas are 0.00372 and 0.00426 for a total alpha = 0.00798 which again yields a conservative test no matter whether the actual value is smaller or larger than the hypothesized value. Finally, we need to ask how these results compare to the asymptotic normal test. The $\alpha = 0.1$ the $\alpha/2$ critical values are ± 1.644853. Each of these values falls on a "flat" region of the score statistic distribution and so that respective tail areas are 0.0247 and 0.0249 giving a total area of 0.0496; thus the test is very conservative compared to 0.1. Similarly, when $\alpha = 0.05$ the asymptotic normal critical values are ± 1.955996, yielding a lower area of 0.00372 and an upper area of 0.0249. Finally, when $\alpha = 0.01$ the normal critical values are ± 2.575827. The lower value is smaller than the smallest possible value of the test statistic (-2.3766 in this case). The upper value has a tail area of 0.0107 which is much larger than the nominal 0.005 desired. Thus we can conclude that critical values chosen by means of the Cornish-Fisher expansion will yield tests much closer to the desired alpha level that the asymptotic normal test when the number of pools is small (m = 300).

Next we repeat the simulation with a different set of 300 pool sizes to see to what extent the results depend on the actual sizes rather than their average properties. In the second simulation, the range of pool sizes was the same but the average size was 37.29 and the variance was 53.39. The results of these calculations are summarized in Table 4.

Table 4: Simulated α values based on 5 replicates assuming $p_0 = 0.0005$, m = 300, and pool sizes between 25 and 50 (with mean pool size 37.29 and variance 53.39).

Requested α	Observed Lower $\alpha/2$	Observed Upper $\alpha/2$
0.1	0.0297**	0.0537
0.05	0.0247	0.0249
0.01	0.0037	0.0042

*Note that the area value with the ** is an approximate value found by simulation. In all other cases, that critical value fell on a "flat" region of the CDF of the score statistic under the null hypothesis and so the probabilities reported are exact.*

Not surprisingly, the results do not change much with small changes in the configuration of pool sizes. Detailed results for more extensive simulations with wider ranges of pool sizes and different values of p_0 are given in Section 4.

3.3 Edgeworth Expansion

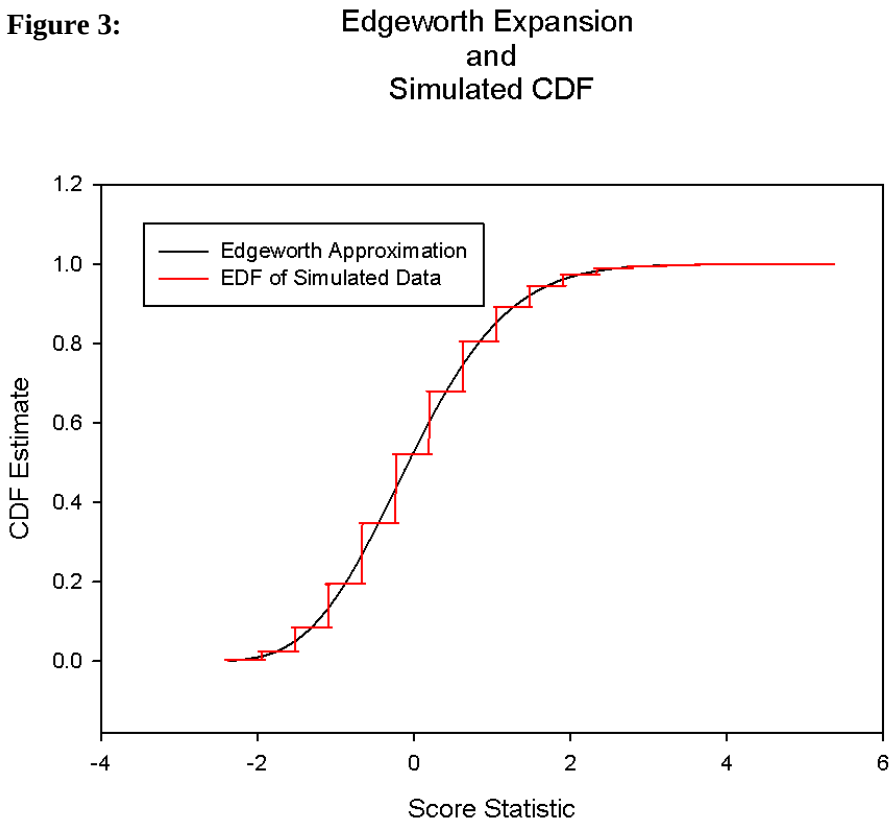
Next we look briefly at using the cumulants to calculate values of the CDF by means of the Edgeworth Expansion. The first few terms of the Edgeworth expansion can be found in Abramowitz and Stegun (1964). Formulas for higher order terms (through the tenth order) are given by Draper and Tierney (1973). Using ten cumulants yields an 8^{th} order formula. The results of applying the formula for the approximate CDF for the data described above (Table 2a) yielded the following plot of the Edgeworth CDF versus the EDF of the statistic based on 200,000 simulation trials (Figure 4 below). The fit is visually reasonable. However, since the Edgeworth approximation is based on a continuous model while the true distribution is discrete, except in rare cases the probabilities calculated with the Edgeworth expansion will be slightly wrong. In some cases they will be too small and in others too large. On the other hand they will still be better than values based on large sample theory. Some probability values calculated by the Edgeworth approximation are compared to values found by simulation in Table 5 for the set of pool sizes and 300 pools used in the tables above.

The entries in Table 5 are in groups of three corresponding to the smallest, middle and largest values in the intervals corresponding to *T = 1,2* and *3* respectively. We see that due to the rapid change in the true CDF over these very short intervals, the values produced by the Edgeworth expansion can be low or high but are about right for the middle of the interval. It is noted in passing, that there did not seem to be any particular advantage in using the Levin transformation. The values calculated in this way agreed to many digits with those found by the truncation method and so since exact values are not known there seems to be not strong reason to do the extra calculations.

Table 5: Approximate CDF using Edgeworth Expansion

Value of Score Statistic, Z_m	Edgeworth Estimate of $P(Z \le Z_m)$	Simulation Estimate of $P(Z \le Z_m)$
-1.94936	0.01187	0.0039
-1.94803	0.01194	0.0131
-1.94669	0.01201	0.0243
-1.52187	0.04988	0.0250
-1.51921	0.05024	0.0843
-1.51653	0.05062	0.0844
-1.09439	0.13345	0.0844
-1.09038	0.13452	0.1274
-1.08637	0.13554	0.1941

Figure 3:



An important application of Edgeworth expansion is the easy calculation of approximate power curves for various alternatives to the null hypothesis. All that is required is the recalculate the cumulants and the Edgeworth expansion, evaluated at the critical point, for each alternative value if p_A of interest.

4. Additional Simulation Results

In this section, we present summary results for additional simulation scenarios. In particular, the speed with which the statistic converges in law to an $N(0,1)$ as measured by comparing the Cornish-Fisher critical values and the asymptotic critical values to "true" values obtained by simulation. In all cases, the simulations are the results of 200,000 trials.

Table 6 summarizes the simulation results for values of α=0.10, 0.05, and 0.01.

Table 6: Simulation results for varying values of α with the pool sizes in the range [25,50] and p_0 = 0.0005.

Case 1: α=0.10

m=number of pools	α=0.10 Critical values	Cornish-Fisher Critical Value	Estimated $\alpha/2$	Simulated 99% CI for critical Value	
1000	Lower	-1.57874	0.04181*	-1.53347	-1.53311
	Upper	1.70712	0.05668*	1.74988	1.75064
3000	Lower	-1.60719	0.04591*	-1.57780	-1.57699
	Upper	1.68122	0.04420	1.67214	1.67306
6000	Lower	-1.61833	0.05273	-1.63822	-1.63696
	Upper	1.67072	0.05037	1.62037	1.71192

Case 2: α=0.05

m=number of pools	α=0.05 Critical values	Cornish-Fisher Critical Value	Estimated $\alpha/2$	Simulated 99% CI for critical Value	
1000	Lower	-1.84869	0.02247	-1.76890	-1.76826
	Upper	2.06302	0.02278	1.98742	1.98853
3000	Lower	-1.89689	0.02388	-1.85016	-1.84834
	Upper	2.02031	0.02529	1.94585	2.07667
6000	Lower	1.91563	0.02711	-1.92613	-1.92484
	Upper	2.002932	0.02343	2.00066	2.00221

Case 3: α=0.01

m=number of pools	α=0.01 Critical values	Cornish-Fisher Critical Value	Estimated $\alpha/2$	Simulated 99% CI for critical Value	
1000	Lower	-2.35109	0.00486*	-2.23891	-2.23744
	Upper	2.77789	0.00459*	2.69128	2.92052
3000	Lower	-2.44958	0.00486*	-2.52276	-2.39002
	Upper	2.69465	0.00507*	2.62162	2.75402
6000	Lower	-2.48736	0.00554	-2.50143	-2.49869
	Upper	2.66059	0.00542	2.66737	2.67279

*Note: Values of $\alpha/2$ with * next to them represent exact values.*

The values in the last three columns of Tables 6 are estimates from a single trial. It is worth noting that the Cornish-Fisher critical values and the 99% confidence intervals for the true critical values are nearer to the N(0,1) critical values (i.e., ±1.64485) for α=0.10 as the number of pools increases. However, for α=0.05 and α=0.01 the Cornish-Fisher critical values and the 99% confidence interval for the true critical values are still not

near the nominal values for the $N(0,1)$ distribution (i.e., ±1.95996 for α=0.05 and ± 2.57583 for α=0.01) even for m=6000.

Next we consider a similar set of simulations where we have expanded the range of pool sizes to the interval [2, 50] and again chosen them to follow a discrete uniform distribution. The results are given in Table 7. Again the calculations are done for $p_0 = 0.0005$ with 200,000 trials in the simulation. The most noteworthy feature in Table 7 is the very poor performance of the Cornish-Fisher expansion when m = 300. The estimated α values for the 3 cases being considered are all significantly lower than the nominal α. This gets worse as α gets smaller.

To understand this, we examine the ranges of values associated with the score statistic for the smaller values of T, the number of positive pools. Table 8 gives this information. For instance, when α=0.10, Table 7 shows that the lower Cornish Fisher critical value is -1.49222. Using the values is Table 8, -1.49222 is between -1.9658 and -1.4526. Thus, the true probability of a score statistic less than or equal to this critical value is equal to $P(T = 0) = 0.0216$. Similarly, the upper Cornish Fisher critical value from Table 7 when α=0.10 is 1.77855 which, when compared to values in Table 8, is between 1.66917 and 2.14049. Thus, the probability that the score statistic is greater than or equal to 1.77855 under the null hypothesis is $1 - P(T \le 7) = 0.03843$.

Another interesting case in Table 7 is when α=0.01 and m=300. Values indicated for the lower critical value reveal a failure for the Cornish Fisher approach and a problem for the simulation case. Table 8 shows that the smallest possible value for the test statistic T was -1.9658012 which is larger than -2.02581, the critical value given by the Cornish Fisher approach. Hence the probability of seeing a value of the test statistic that as small or smaller than -2.02581is clearly zero. The smallest value of the test statistic occurs slightly more than 4330 times in 200,000 trials which is about what you would expect given that $P(T = 0) = 0.021656$. Thus in effect, for small numbers of pools (i.e., too few insects tested) there is effectively no test available with level $\alpha = 0.01$.

The problem is that the number of pools is too small for $p_0 = 0.0005$. When p_0 is this small and the number of pools is 300, the chances of observing no positive pools is about 2% while the chances of seeing one or fewer is 10%. In fact, the probability of finding no positive pools is equal to

$$P(T = 0 \mid p_0; n_1, \cdots, n_m) = (1 - p_0)^{\sum_{j=1}^{m} n_j}$$

Therefore, the problem really has to do with the total number of vector insects tested, not how they were pooled for testing. In general, this number should be large enough that the probability of observing no positive pools is small at the hypothesized value of p_0. In cases where we cannot achieve a large enough number to test, critical values and p-values should almost certainly be calculated by simulation.

Table 7: Simulation results for varying values of α with the pool sizes in the range [2,50] and p_0 = 0.0005.

Case 1: α=0.10

Sample Size m	α=0.10 Critical values	Cornish-Fisher Critical Value	Estimated α/2	Simulated 99% CI for critical Value	
300	Lower	-1.49222	0.02165	-1.44940	-1.44914
	Upper	1.77855	0.03843*	1.65742	1.65794
600	Lower	-1.54024	0.04958*	-1.71481	-1.36011
	Upper	1.74013	0.05037*	1.55208	1.90055
1000	Lower	-1.56450	0.05843*	-1.63636	-1.63586
	Upper	1.71960	0.06103	1.74204	1.74295
3000	Lower	-1.59923	0.04641*	-1.55218	-1.55088
	Upper	1.68859	0.05647	1.69826	1.69968
6000	Lower	-1.61272	0.04723*	-1.58242	-1.58087
	Upper	1.67604	0.04680	1.64409	1.64613

Case 2: α=0.05

Sample Size m	$\alpha = 0.05$ Critical values	Cornish-Fisher Critical Value	Estimated α/2	Simulated 99% CI for critical Value	
300	Lower	-1.69075	0.02199*	-1.45184	-1.45158
	Upper	2.18017	0.01547	2.17241	2.17319
600	Lower	-1.78277	0.01619*	-1.71988	-1.71960
	Upper	2.11730	0.02588*	2.25814	2.26365
1000	Lower	-1.82439	0.02817*	-1.91783	-1.91692
	Upper	2.08359	0.02085*	2.02626	2.02759
3000	Lower	-1.88348	0.02041	-1.87507	-1.87395
	Upper	2.03250	0.02121	2.02387	2.02562
6000	Lower	-1.90621	0.02739	-1.92511	-1.92333
	Upper	2.01174	0.02290	1.98874	1.99129

Case 3: α=0.01

Sample Size m	$\alpha = 0.01$ Critical values	Cornish-Fisher Critical Value	Estimated α/2	Simulated 99% CI for critical Value	
300	Lower	-2.02581	0.00000	-1.96580	-1.96580
	Upper	3.00374	0.00531*	2.69824	3.20227
600	Lower	-2.21256	0.00361*	-2.08343	-2.08279
	Upper	2.88297	0.00577	2.98768	2.99174
1000	Lower	-2.30071	0.00412	-2.20394	-2.20282
	Upper	2.81785	0.00604*	2.86688	2.87004
3000	Lower	-2.42242	0.00462*	-2.52028	-2.36304
	Upper	2.71850	0.00451*	2.67511	2.67980
6000	Lower	-2.46841	0.00573	-2.50175	-2.49871
	Upper	2.67789	0.00511	2.67533	2.68196

*Note: Values of α/2 with * next to them represent exact values.*

Table 8: The distribution values for the random variable T and the associated intervals in which the Score Statistic can have values for the case m = 300, average pool size of 25.5 (pools in the range [2, 50]) and $p_0 = 0.0005$.

T	Smallest Score Statistic	Largest Score Statistic	$P(T = t)$	$P(T \leq t)$
0	-1.9658012	-1.9658012	0.0216563	0.0216563
1	-1.4526098	-1.4464264	0.0836888	0.1053451
2	-0.9394185	-0.9270512	0.1610032	0.2663484
3	-0.4260988	-0.4078061	0.2072204	0.4735688
4	0.0872209	0.1114394	0.1962406	0.6698095
5	0.6005407	0.6306849	0.1484299	0.8182395
6	1.1138604	1.1499305	0.0932740	0.9115134
7	1.6271801	1.6691759	0.0500586	0.9615721
8	2.1404998	2.1884215	0.0234151	0.9849872
9	2.6538195	2.7075377	0.0096956	0.9946828

5. Conclusions

Test based on the Score statistic is a widely-used likelihood-based test procedure. When there is only one parameter to estimate, Score test is easier to implement relative to the other likelihood-based tests (Wald's and likelihood ratio) because there is no need to compute the maximum likelihood estimator for the parameter. However, the Score test procedure relies heavily on its asymptotic normality. In the particular case of pool screening where the probability that a pool will test positive is very small, the number of pools needed for the asymptotic approximation to work well is too large to be practical. Thus, we have found that improvements in the approximations to the distribution and quantiles of the Score Statistic are necessary. Using simulation and the Cornish-Fisher and Edgeworth expansions to approximate the CDF and quantiles of the score statistic are shown to have good performance and are recommended over the conventional asymptotic procedure to be used when performing statistical test about the prevalence in pool screening using the Score statistic. When the prevalence is the number of pools is small for a given value of the prevalence, it is best to use the simulation method.

Appendix

In this appendix we give the formulas for the cumulants $\mathrm{K}_5,\cdots,\mathrm{K}_{10}$ of the Score Statistic for the pool screening model as described above. To reintroduce notation, let $1\le n_1,n_2,\cdots,n_m\le N<\infty$ be m observed pool sizes, let p_0 be the value of the parameter, *p*, in the simple hypothesis test. The random variables $X_k, k=1,\cdots,m$ are independent Bernoulli random variables with probability mass functions,

$$f_{X_j}(x\mid p,n_j)=\left[1-(1-p)^{n_j}\right]^{X_j}\left[(1-p)^{n_j}\right]^{1-X_j},\quad X_j\in\{0,1\},\quad 1\le n_j\le N_{\max}$$

Let $U_m(p\mid X_1,\cdots,X_m;n_1,\cdots,n_m)=\dfrac{1}{(1-p)}\displaystyle\sum_{j=1}^{m}\left\{\frac{n_jX_j}{[1-(1-p)^{n_j}}-n_j\right\}$ and define the Score Statistic as

$$Z_m=\frac{U_m(p_0\mid X_1,\cdots,X_m;n_1,\cdots,n_m)}{\sqrt{Var(U_m(p_0))}}$$

Let $r_k=r_k(p)=\dfrac{n_k(1-p)^{n_k-1}}{[1+(1-p)+\cdots+(1-p)^{n_k-1}]}<1$ for $0<p<1$ and note that $0<r_k\le 1$, $\forall k$.
Finally, define,

$$C=\left\{\sum_{k=1}^{m}\frac{n_k^2(1-p)^{n_k}}{(1-p)^2[1-(1-p)^{n_k}]}\right\}^{1/2}=\left\{\frac{m}{p(1-p)}\left[\frac{1}{m}\sum_{k=1}^{m}n_kr_k\right]\right\}^{1/2}$$

With this notation, it can be shown that,

$$\mathrm{K}_5=\frac{\left(\dfrac{m}{p^4(1-p)}\right)\displaystyle\sum_{k=1}^{m}\frac{1}{m}\left(\frac{n_kr_k^4[-1+14(1-p)^{n_k}-36(1-p)^{2n_k}+24(1-p)^{3n_k}}{(1-p)^{3n_k}}\right)}{C^5}=\bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{3/2}\right)$$

$$\mathrm{K}_6=\frac{\left(\dfrac{m}{p^5(1-p)}\right)\left[\dfrac{1}{m}\displaystyle\sum_{k=1}^{m}\frac{n_kr_k^5[1-30(1-p)^{n_k}+150(1-p)^{2n_k}-240(1-p)^{3n_k}+120(1-p)^{4n_k}]}{(1-p)^{4n_k}}\right]}{C^6}$$

$$=\bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{2}\right)$$

$$\mathrm{K}_7 = \frac{\left(\frac{m}{p^6(1-p)}\right)\left(\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^6 f_7(p,n_k)}{(1-p)^{5n_k}}\right)}{C^7} = \bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{5/2}\right)$$

where

$$f_7(p,n_k) = -1+62(1-p)^{n_k} - 540(1-p)^{2n_k} + 1560(1-p)^{3n_k} - 1800(1-p)^{4n_k} + 720(1-p)^{5n_k}$$

$$\mathrm{K}_8 = \frac{\left(\frac{m}{p^7(1-p)}\right)\left(\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^7 f_8(p,n_k)}{(1-p)^{6n_k}}\right)}{C^8} = \bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{3}\right)$$

where

$$f_8(p,n_k) = 1-126(1-p)^{n_k} + 1806(1-p)^{2n_k} - 8400(1-p)^{3n_k} + 16800(1-p)^{4n_k} - 15120(1-p)^{5n_k} + 5040(1-p)^{6n_k}$$

$$\mathrm{K}_9 = \frac{\left(\frac{m}{p^8(1-p)}\right)\left(\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^8 f_9(p,n_k)}{(1-p)^{7n_k}}\right)}{C^9} = \bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{7/2}\right)$$

where

$$f_9(p,n_k) = -1+254(1-p)^{n_k} - 5796(1-p)^{2n_k} + 40824(1-p)^{3n_k} - 126000(1-p)^{4n_k} + 191520(1-p)^{5n_k} - 141120(1-p)^{6n_k} + 40320(1-p)^{7n_k}$$

and

$$\mathrm{K}_{10} = \frac{\left(\frac{m}{p^9(1-p)}\right)\left(\frac{1}{m}\sum_{k=1}^{m}\frac{n_k r_k^9 f_{10}(p,n_k)}{(1-p)^{8n_k}}\right)}{C^{10}} = \bigcirc\left(\left[\frac{(1-p)}{mp}\right]^{4}\right)$$

where

$$f_{10}(p,n_k) = 1-510(1-p)^{n_k} + 18150(1-p)^{2n_k} - 186480(1-p)^{3n_k} + 834120(1-p)^{4n_k} - 1905120(1-p)^{5n_k} + 2328480(1-p)^{6n_k} - 1451520(1-p)^{7n_k} + 326880(1-p)^{8n_k}$$

References:

Abramowitz, M. and Stegun, I., "Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables", National Bureau of Standards Applied Mathematics Series, 55, 1964.

Barker, J.T., "Statistical Estimates of Infection Potential Based on PCR Pool Screening with Unequal Pool Sizes", PhD thesis, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 2000

Chiang, C., and Reeves, W., "Statistical Estimation of Virus Infection Rates in Mosquito Vector Populations", Am J Hyg, v. 75, p. 377-391, 1962.

Draper, N.R. and Tierney, D.E.," Exact formulas for additional terms in some important series expansions", Communications in Statistics, Vol1, No. 6, pp495-524 (1973).

Fessler, T and Ford,W, "HURRY: An Acceleration Algorithm for Scalar Sequences and series", ACM Transactions on Mathematical Software, Vol 9, No. 3, pp 346-354 (1983).

Field, C. and Ronchetti, E., "Small Sample Asymptotics", Institute of Mathematical Statistics, Lecture Notes-Monographs Series, Vol 13, Hayward California, 1990.

Frydman, H. and Simon, G, "Discrete Quantile Estimation", NYU Faculty Archive, Report SOR-2007-2, http://hdl.handle.net/2451/26295, pp 1-21, 2007.

Gibbons, J.D. and Chakraborti, S., "Nonparametric Statistical Inference: Fourth Edition", Statistics: Textbooks and Monographs, Vol 168, Marcel Dekker, Inc., New York, 1993.

Hepworth G., "Exact Confidence Intervals for Proportions Estimated by Group Testing", Biometrics, v. 52, p. 1134-1146, 1996.

Katholi, C., Toe, L., Merriweather A., and Unnasch, T., "Determining the Prevalence of Onchocerca *Volvulus* Infection in Vector Populations by Polymerase Chain Reaction Screening of Pools of Black Flies", J. Infect Dis., v. 172, p. 1414-1417, 1995

Katholi C.R., and Unnasch, T.R., "Important Experimental Parameters for Determining Infection Rates in Arthropod Vectors Using Pool Screening Approaches", Am J Trop Med Hyg, v. 74, p. 779-785, 2006.

Koehler,E, Brown, E. and Haneuse, S., "On the assessment of Monte Carlo Errors in Simulation-Based Statistical Analyses", The American Statistician, Vol 63, No. 2, pp155-162, 2009.

Lee, Yoong-Sin and Lee, Moy Chee,"On the Derivation and Computation of the Cornish-Fisher Expansion", Australian Journal of Statistics, 34(3), pp 443-450, 1992.

Lee, Yoong-Sin and Lin, Ting-Kwong, “High Order Cornish –Fisher Expansion”, Algorithm 269, Applied Statistics, 41(1), pp 233-240, 1992.

Scheffe’, H. and Tukey, J.,”Non-parametric estimation I. Validation of Order Statistics”, Annals of Mathematical Statistics, Vol 16, pp 187-192, 1945.

Sen, P.K. and Singer, J.M., “Large Sample Methods in Statistics”, Texts in Statistical Science, Chapman&Hall/CRC, 1993.

Smith, D.A. and Ford, W.F., “Acceleration of Linear and Logarithmic Convergence”, SIAM Journal of Numerical Analysis, Vol 16, No. 2, pp 223-240,1979.

Smith, D.A. and Ford, W.T, Numerical Comparison of Nonlinear Convergence Accelerators”, Vol 38, No. 158, pp 481-499, 1982.

Tebbs, J.M. and McCann, M.H., “Large-Sample Hypothesis Tests for Stratefied Group-Testing Data”, Journal of Agricultural, Biological and Environmental Statistics, Vol 12, No. 4, pp534-551, 2007.