Biostatistics Department Technical Report

BST2011-001

Sample Size Calculation in Pool Screening

Charles R. Katholi, Ph.D
Inmaculada Aban, Ph.D.
July 2011

Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
ckatholi@uab.edu
Sample Size Calculation in Pool Screening

For pool screening (group testing) the sampling model is that of a series of Bernoulli trials and the observed number of positive pools then has a Binomial distribution. Suppose the we collect pools of size \( n \) and that the underlying prevalence we wish to test is denoted by \( p_0 \). If we collect \( m \) pools and denote the number of positive pools by \( T \) then the probability mass function of \( T \) is

\[
P(T = t \mid p_0, n, m) = \binom{m}{t} \left[ 1 - (1 - p_0)^n \right]^t \left[ (1 - p_0)^n \right]^{m-t}
\]

A hypothesis test of \( H_0 : p = p_0 \) versus \( H_a : p < p_0 \) can be based on \( T \) and it can be shown that it is the uniformly most powerful test of this kind. The test has the rejection region \( T \leq T_c \) where \( T_c \) is chosen so that \( P(T \leq T_c \mid p = p_0) = \alpha \).

When \( p_0 \) is small (say 25 per 10000), there is a minimal number, \( m_0 \), of pools of size \( n \) required for the test to even be meaningful. To see this, we note that the probability of observing \( T = 0 \) is

\[
P(T = 0 \mid p_0, n, m) = (1 - p_0)^{mn}
\]

and that in order to make a test this probability must be at least as small as the proposed \( \alpha \) level of the test for otherwise observing no positive pools would not reject the null hypothesis. In what follows, we shall make the rule that the minimum number of pools collected, \( m_0 \) must be large enough so that

\[
P(T \leq 1 \mid p_0, n, m_0) \leq \alpha \text{ and } P(T \leq 2 \mid p_0, n, m_0) > \alpha \quad (1.1)
\]

Because the distribution of \( T \) is discrete, we will have to settle for a level \( \alpha \) test rather than a size \( \alpha \) test.

Finding a value of \( m_0 \) such that \( P(T = 0) \leq \alpha \) is very straightforward. As noted previously, \( P(T = 0 \mid p_0, n, m) = (1 - p_0)^{mn} \) and this will be less than \( \alpha \) for any fixed \( p_0 \)

and \( n \) if \( m_0' \geq \left\lceil \frac{\ln(\alpha)}{n \ln(1 - p_0)} \right\rceil + 1 \) where \( \lceil y \rceil \) stands for the greatest integer less than \( y \).

Next, since the conditions of equation (1.1) can be viewed as the requirement that

\[
P(T \leq 1 \mid p_0, n, m) = (1 - p)^{mn} + m \left[ 1 - (1 - p)^n \right] \left[ (1 - p)^n \right]^{m-1} \approx \alpha
\]
The desired value of $m_0$ can then be found by solving the inequality

$$-\left[ (m-1)n \ln(1-p_0) + \ln \left( \frac{1 - (1-p_0)^n}{m - (m-1)(1-p_0)^n} \right) \right] \geq -\ln(\alpha)$$

(1.2)

subject to the constraint that $m \geq m'_0$. Note that the derivative (with respect to $m$) of the function on the left hand side of this equation is equal to

$$-n \ln(1-p_0) - \frac{\left[ 1 - (1-p_0)^n \right]}{m - (m-1)(1-p_0)^n}$$

The first term is positive by virtue of the fact that $0 < p_0 < 1$. The second term requires some discussion. It is clear that the quantity

$$\frac{\left[ 1 - (1-p_0)^n \right]}{m - (m-1)(1-p_0)^n} = p_0 \frac{1 + (1-p_0) + (1-p_0)^2 + \cdots + (1-p_0)^{n-1}}{m - (m-1)(1-p_0)^n}$$

is positive for all $m, n \geq 1$ and $0 < p_0 < 1$. The numerator is bounded above by $np_0$ and the denominator takes on its smallest value when $m = 1$. Hence it follows that this quantity is bound above by $np_0$ for all $m \geq 1$ and decreases as $m$ gets large. Hence,

$$-\frac{\left[ 1 - (1-p_0)^n \right]}{m - (m-1)(1-p_0)^n} \geq -np_0$$

for all $m, n \geq 1$ and $0 < p_0 < 1$. Furthermore,

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

so that

$$-n \ln(1-p_0) - \frac{\left[ 1 - (1-p_0)^n \right]}{m - (m-1)(1-p_0)^n} \geq np_0 + np_0 \left[ \frac{1}{2} p_0 + \frac{1}{3} p_0^2 + \cdots \right] - np_0 > 0$$

Thus, the derivative is positive for all $m \geq 1$ and so the function on the left hand side of equation (1.2) is strictly increasing as a function of $m$. In addition, since the second derivative of the left hand side of equation (1.2) with respect to $m$ is always negative the function is concave downward. These conditions are generally sufficient to guaranteed the stability of Newton’s method as a method for generating trial solutions for solving the
inequality. Newton’s method will lead, in general, to a non-integer solution and so the
final result should be the nearest integer larger than the observed solution.

Computation of the sample size required to achieve a given power against a
specific alternative is more complex. In this case two inequalities must be satisfied
simultaneously; one that identifies the critical value for the test of level \( \alpha \) and a second
one which specifies the power. Thus, for any sample size \( m \), level \( \alpha \) and power \( \beta \) we
must first find the critical value \( T_c \) such that the rejection region for an alternative of the
form \( H_A : p < p_0 \) is \( T \leq T_c \). This means that \( T_c \) satisfies the conditions

\[
P(T \leq T_c \mid m, n, p_0) \leq \alpha \quad \text{and} \quad P(T \leq T_c + 1 \mid m, n, p_0) > \alpha
\]

Once \( T_c \) is found, then by definition

\[
\beta = \beta_m = P(T \leq T_c \mid m, n, \theta_A)
\]

The value of \( m \) is varied until \( \beta_m \geq \beta_{\text{desired}} \) while \( \beta_{m-1} < \beta_{\text{desired}} \). Since, in general, the
power function is increasing as a function of the sample size, this condition insures that
the sample size is the minimum necessary.

This process can be greatly facilitated if a good initial estimate of \( m \) can be
obtained. Asymptotic test methods provide a way to find such an estimate. From
Lehmann (2010) an asymptotic test for the parameter \( \theta_0 \) of a Binomial random variable \( T 
\)
with success probability \( \theta_0 \) and sample size \( m \), is based on the test statistic,

\[
Z = \sqrt{m} \frac{\left( \frac{T}{m} - \theta_0 \right)}{\sqrt{\theta_0(1-\theta_0)}}
\]

which is asymptotically \( N(0,1) \). Consideration is given to a sequence of alternatives of
the form \( \theta_A = \theta_0 + \frac{\Delta}{\sqrt{m}} \) where \( \Delta \) is positive or negative depending on the direction of
the one-sided test. This leads to the sample size calculation formula

\[
m = \frac{(u_\alpha - u_\beta)^2}{\Delta^2 \theta_0(1-\theta_0)}
\]

where \( u_\alpha \) and \( u_\beta \) are appropriate critical values from the \( N(0,1) \) distribution.

For the pool screening model, \( \theta_0 = 1 - (1 - p_0)^\tau \), and for an alternative of the form
\( \theta_A = \theta_0 + \frac{\Delta}{\sqrt{m}} \) we lead to a \( \theta_A(\Delta) = \theta_0 + \frac{\Delta}{\sqrt{m}} \) where by expanding
\[ \theta_4(\Delta) = 1 - (1 - p_a)^n = 1 - \left(1 - \left[p_0 + \frac{\Delta}{\sqrt{n}}\right]\right)^n \]
in a Taylor expansion about \( \Delta = 0 \) and dropping all but the first order terms we obtain the approximation \( \tilde{\Delta} \cong n(1 - p_0)^{n-1} \Delta \). This leads to the formula

\[
m \cong \frac{(u_\alpha - u_\beta)^2}{(n(1 - p_0)^{n-1} \Delta)^2} \left[1 - (1 - p_0)^n\right]\left[(1 - p_0)^n\right] \tag{1.3}
\]

This formula is simple and can be easily applied with a hand calculator and a table for the standard normal distribution to find the quantities \( u_\alpha \) and \( u_\beta \). It would be of interest to see if these values can be used to find the sample size for the exact test we are investigating. If it could, this would save a great deal of complex computation. To this end we consider the results given in Tables 1 and 2 which involve using the asymptotic sample size in making the test. Thus the results indicated represent the quantities associated with basing the test on the asymptotic sample size (asympt.) and for comparison purposes basing the test on the smallest \( m \) such that

\[
P(T \leq T_c \mid m, n, p_0) \leq \alpha \quad \text{and} \quad P(T \leq T_c + 1 \mid m, n, p_0) > \alpha
\]

and \( \beta_m \geq \beta_{\text{desired}} \) while \( \beta_{m-1} < \beta_{\text{desired}} \) where \( T_c \) is the critical value for the test.

Table 1 below gives example results utilizing this formula for the case of pools of size \( n = 50 \) and target values of \( \alpha = 0.05 \) and \( \beta = 0.80 \) and test \( H_0 : p = p_0 \) vs \( H_1 : p < p_0 \).

<table>
<thead>
<tr>
<th>( p_A )</th>
<th>( 0.8p_0 )</th>
<th>( 0.6p_0 )</th>
<th>( 0.4p_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_0 )</td>
<td>Asympt.</td>
<td>Minimal</td>
<td>Asympt.</td>
</tr>
<tr>
<td>25</td>
<td>1312</td>
<td>1196</td>
<td>328</td>
</tr>
<tr>
<td>10,000</td>
<td>134</td>
<td>122</td>
<td>28</td>
</tr>
<tr>
<td>( T_c )</td>
<td>0.0425</td>
<td>0.0491</td>
<td>0.0376</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.8156</td>
<td>0.8015</td>
<td>0.8161</td>
</tr>
<tr>
<td>5</td>
<td>697</td>
<td>634</td>
<td>174</td>
</tr>
<tr>
<td>1,000</td>
<td>136</td>
<td>123</td>
<td>29</td>
</tr>
<tr>
<td>( T_c )</td>
<td>0.0486</td>
<td>0.0499</td>
<td>0.0455</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.8354</td>
<td>0.8064</td>
<td>0.8729</td>
</tr>
<tr>
<td>1</td>
<td>395</td>
<td>348</td>
<td>99</td>
</tr>
<tr>
<td>100</td>
<td>139</td>
<td>122</td>
<td>30</td>
</tr>
<tr>
<td>( T_c )</td>
<td>0.0403</td>
<td>0.0497</td>
<td>0.0369</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.8189</td>
<td>0.8008</td>
<td>0.8624</td>
</tr>
</tbody>
</table>

Table 1
The column labeled “Asympt.” corresponds to basing the test on the asymptotic sample size while the one labeled “Minimal” corresponds to finding the sample size by the search method described above. The table indicates that at least for these few examples, the sample size estimated using the asymptotic formula is too large. It is also apparent that choosing the smallest sample size to achieve the desired power often leads to a critical value with associated alpha level considerably smaller than the target level. This is simply a result of the fact that the test is based on a discrete distribution.

Similar calculations for the test \( H_0 : p = p_0 \) versus \( H_A : p > p_0 \) are given in table 2. Again the column labeled “Asympt.” corresponds to basing the test on the asymptotic sample size while the one labeled “Minimal” corresponds to finding the sample size by searching until the criteria described above are met. In this case, the asymptotic sample size is generally too small to achieve that desired power.

<table>
<thead>
<tr>
<th>( p_A )</th>
<th>( 1.2 p_0 )</th>
<th>( 1.4 p_0 )</th>
<th>( 1.6 p_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_0 )</td>
<td>Asympt.</td>
<td>Minimal</td>
<td>Asympt.</td>
</tr>
<tr>
<td>25</td>
<td>( m )</td>
<td>1312</td>
<td>1446</td>
</tr>
<tr>
<td></td>
<td>( T_C )</td>
<td>175</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>0.0439</td>
<td>0.0498</td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>0.7950</td>
<td>0.80161</td>
</tr>
<tr>
<td>5</td>
<td>( m )</td>
<td>697</td>
<td>777</td>
</tr>
<tr>
<td></td>
<td>( T_C )</td>
<td>174</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>0.0431</td>
<td>0.0496</td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>0.7434</td>
<td>0.8021</td>
</tr>
<tr>
<td>1</td>
<td>( m )</td>
<td>396</td>
<td>449</td>
</tr>
<tr>
<td></td>
<td>( T_C )</td>
<td>173</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>0.0497</td>
<td>0.0495</td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>0.7585</td>
<td>0.8045</td>
</tr>
</tbody>
</table>

Table 2

The construction of an algorithm to find the minimum number of pools as just described can be simplified if we make use of the well known relationship between the upper tail probabilities of the binomial distribution and the incomplete beta function. We recall the following identities (Abramowitz and Stegun):

\[
\sum_{s=0}^{m} \binom{m}{s} \theta^s (1- \theta)^{m-s} = I_\theta(a,m-a+1)
\]

(1.4)

where
\[ I_0(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \int_0^\beta t^{\alpha-1}(1-t)^{\beta-1} \, dt \]  

(1.5)

and

\[ I_0(\alpha, \beta) = 1 - I_0(\beta, \alpha) \]  

(1.6)

Combining equations (1.4) and (1.6) leads to the computational formula

\[ \sum_{s=0}^{a} \binom{m}{s} \theta^s (1-\theta)^{m-s} = 1 - \sum_{s=\alpha+1}^{m} \binom{m}{s} \theta^s (1-\theta)^{m-s} = I_{1-o}(m - a, a + 1) \]  

(1.7)

Because the incomplete beta function is continuous in its parameters and if \( a \) is not restricted to integers, it is possible to find an unique number \( \tilde{a} \) such that for any fixed integer \( m \), \( I_{1-o}(m - \tilde{a}, \tilde{a} + 1) = \alpha \). Next note that as \( m \) is increased in unit increments 

(i.e. \( m \to m+1 \) ) , \( \tilde{a} \) changes slowly when \( m \) is large and so \( \lfloor \tilde{a} \rfloor \), the integer portion of \( \tilde{a} \) remains constant over a range of values of \( m \). Thus, since the critical value of the test statistic is an integer, the critical value remains the same over a range of \( m \) values. Obviously, since \( \tilde{a} \) is not an integer we expect that the \( \alpha \) level of the test based on \( \lfloor \tilde{a} \rfloor \) will be different from the value \( \alpha \) desired. The fact that it will be less than \( \alpha \) follows from the fact that the incomplete beta function is an increasing function of \( a \) when \( m \) is fixed and a decreasing function of \( m \) when \( a \) is fixed (Gun, 1965). For integer increments this is well known. In particular, note the following identities for the incomplete beta function (Abramowitz and Stegun, page 944):

\[ I_+ (r, s) = I_+ (r + 1, s) + \frac{\Gamma(r + s)}{\Gamma(r + 1)\Gamma(s)} x^r (1-x)^s \]  

(1.8)

\[ I_+ (r, s) = I_+ (r - 1, s + 1) - \frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s + 1)} x^{r-1} (1-x)^s \]  

(1.9)

where \( r \) and \( s \) need not be integers as long as \( r - 1 > 0 \). From these it follows immediately that \( I_+ (r, s) > I_+ (r + 1, s) \) and \( I_+ (r, s) < I_+ (r - 1, s + 1) \). Gun, (1965) showed that for \( 0 < x < 1 \), \( I_+ (r, s) \) is a monotonically decreasing function of \( r \) and a monotonically increasing function of \( s \). His approach leads to the following result.

**Lemma**: Let \( \varepsilon > 0 \) be any positive quantity then \( I_+ (r, s) < I_+ (r - \varepsilon, r + \varepsilon) \)

Proof: Consider the quantity
\[ h = B(r, s)B(r - \varepsilon, r + \varepsilon) \left[ I_1(r, s) - I_1(r - \varepsilon, r + \varepsilon) \right] \]

where \( B(\bullet, \bullet) \) is the complete beta function. Since \( B(\bullet, \bullet) \) is positive whenever the arguments are positive, the algebraic sign of \( h \) depends on the difference in the two incomplete beta functions. We shall show that \( h < 0 \) when \( \varepsilon > 0 \). To this end, note that

\[
\begin{align*}
  h &= \int_0^1 z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} \, dz \int_0^y (1-y)^{r-1} \, dy - \int_0^1 z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} \, dz \int_0^y (1-y)^{r-1} \, dy \\
  &= \int_0^y \int_0^1 z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} y^{r-1} (1-y)^{r-1} \, dy \, dz - \int_0^1 \int_0^y z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} y^{r-1} (1-y)^{r-1} \, dy \, dz
\end{align*}
\]

If the outer integral is divided into two segments, \((0, x)\) and \((x, 1)\) in each case, the terms involving the interval segment \((0, x)\) from the two integrals are equal and hence cancel. This we have that,

\[
(1.10)
\]

\[
\begin{align*}
  h &= \int_x^1 \int_0^y z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} y^{r-1} (1-y)^{r-1} \, dy \, dz - \int_0^x \int_0^y z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} y^{r-1} (1-y)^{r-1} \, dy \, dz \\
  &= \int_x^1 \int_0^y z^{r-\varepsilon-1} (1-z)^{s+\varepsilon-1} y^{r-1} (1-y)^{r-1} \left[ \left( \frac{1-z}{z} \right)^\varepsilon - \left( \frac{1-y}{y} \right)^\varepsilon \right] \, dy \, dz
\end{align*}
\]

Noting that the double integral is over the region \( \mathbb{R} = \{(y, z) | 0 < y < x, x < z < 1\} \) so that \( y < z \) leads to the fact that \( \left( \frac{1-y}{y} \right) > \left( \frac{1-z}{z} \right) > 0 \). Thus for any \( \varepsilon > 0 \),

\[
\left( \frac{1-z}{z} \right)^\varepsilon - \left( \frac{1-y}{y} \right)^\varepsilon < 0
\]

so the integrand defining \( h \) is always negative and thus \( h < 0 \) which completes the proof.

To apply these here note that \( r = m - a \) and \( s = a + 1 \), and that the value of \( T_c = T_c(m, p_0) \), the critical value for the test, is \( \left\lceil \tilde{a} \right\rceil \), the integer part of \( \tilde{a} \). Thus, if \( \tilde{a} \) is such that \( I_{1,0}(m - \tilde{a}, \tilde{a} + 1) = \alpha \) then \( I_{1,0}(m - T_c, T_c + 1) \leq \alpha \). Furthermore, since \( I_*(r, s) > I_*(r+1, s) \), a unit increase in \( m \) requires a compensatory increase in \( \tilde{a} \) if \( I_{1,0}(m - \tilde{a}, \tilde{a} + 1) \) is to equal \( \alpha \). This increase will generally be less than one and so if we denote this new value of \( \tilde{a} \) by \( \tilde{a}' \) it will be the case for a range of values of \( m \) that \( T_c(m, p_0) = T_c(m + 1, p_0) \) since \( \left\lfloor \tilde{a} \right\rfloor = \left\lfloor \tilde{a}' \right\rfloor \). This is illustrated clearly in Table 3.
that for \( m \in [1187, 1195] \), \( \alpha \) is increasing while the value of \( T_c \) remains constant at \( T_c = 121 \). As \( m \) increases, there will be a point where \( T_c(m + k, p_0) = T_c(m, p_0) + 1 \); that is, where the critical value for the test takes the value of the next integer. It is at this value of \( m \) that the test has an alpha level closest to the desired value since over the range of values of \( m \) leading to the same critical value, \( I_{1-\theta}(m - T_c, T_c + 1) \) is a decreasing function of \( m \). These observations are illustrated by the example in Table 3.

As noted previously, in searching for the sample size, \( m \), we need to find a size so that both the Type I error and the power objectives are simultaneously met. For any alternative, increasing the sample size will should increase the power, but again because of the discrete nature of the quantities involved, the power also decreases as the critical value is constant over a range of values of \( m \). Thus referring to Table 3 we observe this behavior on the intervals \( 1187 \leq m \leq 1196 \) and \( 1197 \leq m \leq 1204 \).

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \alpha ) ( \in I_{1-\theta}(m - \tilde{\alpha}, \tilde{\alpha} + 1) = 0.05 )</th>
<th>( T_c ) ( = \lceil \tilde{\alpha} \rfloor )</th>
<th>( \alpha )</th>
<th>power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1185</td>
<td>120.8876</td>
<td>120</td>
<td>0.0420</td>
<td>0.7765</td>
</tr>
<tr>
<td>1186</td>
<td>120.9975</td>
<td>120</td>
<td>0.0411</td>
<td>0.7736</td>
</tr>
<tr>
<td>1187</td>
<td>121.1075</td>
<td>121</td>
<td>0.0490</td>
<td>0.7990</td>
</tr>
<tr>
<td>1188</td>
<td>121.2174</td>
<td>121</td>
<td>0.0479</td>
<td>0.7963</td>
</tr>
<tr>
<td>1189</td>
<td>121.3274</td>
<td>121</td>
<td>0.0469</td>
<td>0.7936</td>
</tr>
<tr>
<td>1190</td>
<td>121.4373</td>
<td>121</td>
<td>0.0459</td>
<td>0.7909</td>
</tr>
<tr>
<td>1191</td>
<td>121.5473</td>
<td>121</td>
<td>0.0449</td>
<td>0.7881</td>
</tr>
<tr>
<td>1192</td>
<td>121.6572</td>
<td>121</td>
<td>0.0440</td>
<td>0.7853</td>
</tr>
<tr>
<td>1193</td>
<td>121.7672</td>
<td>121</td>
<td>0.0430</td>
<td>0.7825</td>
</tr>
<tr>
<td>1194</td>
<td>121.8772</td>
<td>121</td>
<td>0.0421</td>
<td>0.7797</td>
</tr>
<tr>
<td>1195</td>
<td>121.987</td>
<td>121</td>
<td>0.0412</td>
<td>0.7769</td>
</tr>
<tr>
<td>1196</td>
<td>122.097</td>
<td>122</td>
<td>0.0491</td>
<td>0.8020</td>
</tr>
<tr>
<td>1197</td>
<td>122.2071</td>
<td>122</td>
<td>0.0480</td>
<td>0.7993</td>
</tr>
<tr>
<td>1198</td>
<td>122.3171</td>
<td>122</td>
<td>0.0470</td>
<td>0.7966</td>
</tr>
<tr>
<td>1199</td>
<td>122.4271</td>
<td>122</td>
<td>0.0460</td>
<td>0.7939</td>
</tr>
<tr>
<td>1200</td>
<td>122.537</td>
<td>122</td>
<td>0.0450</td>
<td>0.7912</td>
</tr>
<tr>
<td>1201</td>
<td>122.647</td>
<td>122</td>
<td>0.0441</td>
<td>0.7885</td>
</tr>
<tr>
<td>1202</td>
<td>122.757</td>
<td>122</td>
<td>0.0431</td>
<td>0.7857</td>
</tr>
<tr>
<td>1203</td>
<td>122.867</td>
<td>122</td>
<td>0.0422</td>
<td>0.7830</td>
</tr>
<tr>
<td>1204</td>
<td>122.977</td>
<td>122</td>
<td>0.0413</td>
<td>0.7802</td>
</tr>
<tr>
<td>1205</td>
<td>123.087</td>
<td>123</td>
<td>0.0492</td>
<td>0.8049</td>
</tr>
<tr>
<td>1206</td>
<td>123.197</td>
<td>123</td>
<td>0.0481</td>
<td>0.8023</td>
</tr>
</tbody>
</table>

Note: This table was constructed with \( p_0 = 0.0025 \), \( p_0 = 0.0020 \), \( \alpha = 0.05 \) and the desired power = 0.80.

Table 3
Thus, the search requires that we find the first value of $m$ for which all conditions are met. The behavior of the type I error and power on intervals of $m$, with constant critical value preclude the use of sophisticated search algorithms and so a simple strategy of starting at the minimum value of the sample size for which the test is properly defined and incrementing $m$ by one at each step until the conditions are met, seems to be the only reasonable computational strategy. The calculation of the asymptotic sample size is helpful in this effort, particularly for the hypothesis test

$$H_0: p = p_0 \quad \text{versus} \quad H_A: p > p_0$$  \hspace{1cm} (1.11)

As noted in Table 2, illustrative calculations suggest that in this case, the asymptotic sample size is generally smaller than needed, so the search can be started at this value rather than with the value of $m_0$ defined above. The computational approach associated with the calculation of the sample size for the hypothesis test

$$H_0: p = p_0 \quad \text{versus} \quad H_A: p < p_0$$  \hspace{1cm} (1.12)

can be summarized in the following pseudo code. The code for the case of equation (1.11) is similar.

**Algorithm:** Given $p_0, p_a$ and $n$, let $\theta_0 = 1 - (1 - p_0)^n$ and $\theta_a = 1 - (1 - p_a)^n$

1. Calculate $m_0$ by solving the inequality (1.2) finding trial solutions by Newton’s method.

2. Calculate $m_a$ the asymptotic sample size based on equation (1.3).

3. Calculate $T_c(m_a)$ and find the starting value of $m$ for the search as follows:
   
   (i). Solve the equation $I_{1 - \theta_0}(m_a - \tilde{a}, \tilde{a} + 1) = \alpha$ (defined in equation (1.7)) for $\tilde{a}$ then calculate $T_c(m_a) = \lfloor \tilde{a} \rfloor$.

   (ii). Calculate $\beta = power = P(T \leq T_c(m_a)) = I_{1 - \theta_a}(m - T_c, T_c + 1)$

   (iii). If $\beta > \beta_{desired}$, set $m = m_0$ else set $m = m_a$.

4. Search for the smallest $m$ which meets the criteria set out previously.

(i). Set $m = m + 1$
(ii). Solve the equation $I_{1-a_0}(m - \tilde{a}, \tilde{a} + 1) = \alpha$ (defined in equation (1.7)) for $\tilde{a}$ then calculate $T_c(m) = \lceil \tilde{a} \rceil$.

(iii). Calculate $\beta = power = P(T \leq T_c(m_a)) = I_{1-a_0}(m - T_c, T_c + 1)$

(iv). If $\beta < \beta_{desired}$ repeat steps 4.(i) through 4.(iii); else exit with sample size equal to $m$.

End of algorithm

The pseudo code for the test of equation (1.11) is similar. The main difference is in step 3. where the starting value of the search is defined and in the need to set an upper bound on the practical sample size if the search is initiated starting at $m_a$. In step 3, if $\beta(m_a) > \beta_{desired}$ then the search “back tracks” by halving $m$, until $\beta(m) < \beta_{desired}$ and then the search proceeds as in step 4.

The calculations described in this report are implemented in a Windows program called PS_SampleSize which is available from the authors upon request.

References:

