Technical Report

From: Mark Wolters
To: file
Date: 10-Jun-2007
Subject: **Sample Size for Estimating Multinomial Proportions**

Problem Statement

A total of N patients will be classified into groups A, B, C, and D simultaneously by each of two devices. The first device is the gold standard; its outcome will be called the *true state*. The second device is a new technology; its outcome will be called the *measured state*. The goal of the study is to assess the performance of the new device relative to the standard.

The four categories (A, B, C, D) are disease states, with state A being normal. The true state of any patient in the study will not be known with certainty before the test. Let the outcome of any trial be denoted by a two-letter combination, with the first character indicating the true state, and the second character indicating the measured state. So, for example, outcome AB indicates a person who was truly normal but was measured to be in disease state B. Note that there are 16 possible outcomes for any individual:

AA, BB, CC, DD, AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC.

These outcomes have been ordered with the correct classifications (AA, BB, CC, DD) first; if the new device performs well, the large majority of the trials should result in one of these four outcomes. The remaining 12 classifications constitute different types of errors.

From a statistical perspective, each patient may be viewed as having a prior probability π_i of eventually being classified into outcome i. That is,

P(a subject's test results are AA) = π_{AA}
P(a subject's test results are BB) = π_{BB}
P(a subject's test results are CC) = $\pi_{CC,}$

and so on for the 16 possible outcomes.

In this framework, the goal of the experiment is to estimate the classification probabilities, π_i. Viewing each category individually, each π_i is a binomial proportion. Viewing all 16 categories together, the vector $\boldsymbol{\pi} = (\pi_{AA}, \pi_{BB}, \ldots, \pi_{DC})$ is a set of multinomial probabilities[1] such that $\Sigma\pi_i = 1$.

[1] The data from this experiment—the number of counts in each of the 16 outcomes—may be viewed as a sample from a multinomial distribution with N trials and probability vector $\boldsymbol{\pi}$. Multinomial distributions arise when independent, identical trials are repeated with fixed probabilities of more than two outcomes (e.g., the results of 100 rolls of a die is a sample from a multinomial distribution with N = 100 and π_i = 1/6, i = 1, 2, … 6). The number of outcomes in each specific category has a binomial distribution.

Based on this description of the problem, the following objectives can be stated for the statistical analysis to follow:

1. Describe a method for estimating the π_i's, and give formulas for calculating the uncertainty associate with the estimate.
2. Provide methods for determining the sample size N, such that the multinomial probabilities $\boldsymbol{\pi}$ may be estimated with sufficient precision.
3. Use the above methods to analyze the experiment and make recommendations on sample size and analysis methods.

These objectives will be considered in the following sections. The starting point is to review concepts related to confidence intervals for binomial proportions.

Background

The estimate of each π_i will be called p_i. Each p_i can be calculated straightforwardly; it is just the proportion of the N subjects that had outcome i. Let n_i be the number of subjects with outcome i. Then

$$p_i = \frac{n_i}{N} . \qquad (1)$$

Once the 16 p_i's have been calculated, the performance of the new device can be assessed in whatever ways are considered appropriate. For presentation purposes, it is probably useful to show the results of the experiment in the form of a 4x4 contingency table, as in Table 1.

Table 1. A display of results from the experiment.

a) as counts

True \ Measured	A	B	C	D	Sum
A	n_{AA}	n_{AB}	n_{AC}	n_{AD}	$n_{A.}$
B	n_{BA}	n_{BB}	n_{BC}	n_{BD}	$n_{B.}$
C	n_{CA}	n_{CB}	n_{CC}	n_{CD}	$n_{C.}$
D	n_{DA}	n_{DB}	n_{DC}	n_{DD}	$n_{D.}$
	$n_{.A}$	$n_{.B}$	$n_{.C}$	$n_{.D}$	

b) as proportions

True \ Measured	A	B	C	D	Sum
A	p_{AA}	p_{AB}	p_{AC}	p_{AD}	$p_{A.}$
B	p_{BA}	p_{BB}	p_{BC}	p_{BD}	$p_{B.}$
C	p_{CA}	p_{CB}	p_{CC}	p_{CD}	$p_{C.}$
D	p_{DA}	p_{DB}	p_{DC}	p_{DD}	$p_{D.}$
	$p_{.A}$	$p_{.B}$	$p_{.C}$	$p_{.D}$	

In the above tables, the "dot notation" has been used to represent the row and column totals (e.g. $p_{A.}$ is the sum of the proportions of subjects with true state A). Once the table has been constructed, any other desired probabilities can be calculated. For example:

- Unconditional probabilities can be read directly:
  P(truth is A and measured is C) = p_{AC}
- Conditional probabilities can be calculated easily:
  P(measure C, given patient is A) = $p_{AC}/p_{A.}$

- A collapsed 2x2 table can be constructed. This could be useful, for example, if one wants to speak in the familiar terms of sensitivity and specificity for a particular disease state. Consider, for example, state A:

		Measured A	**Measured** not A
True A		p_{AA}	$p_{AB}+p_{AC}+p_{AD}$
True not A		$p_{BA}+p_{CA}+p_{DA}$	$p_{BB}+p_{BC}+p_{BD}+p_{CB}+p_{CC}+p_{CD}+p_{DB}+p_{DC}+p_{DD}$

Confidence Interval for a Proportion

The above points illustrate that the proportions p_i, as a group, constitute a thorough summary of the performance of the new device relative to the gold standard. The simple calculations involving the p_i's are of little use, however, if the proportions are estimated with insufficient precision. The point of studying the effect of sample size is to ensure that the estimates p_i are sufficiently precise to permit the new device's performance to be assessed with a reasonable level of certainty.

One measure of precision for a statistical estimate is the confidence interval. For a proportion such as p_i, the usual interval used to provide reasonable bounds is calculated as follows:

$$\left(p_i - d_i,\ p_i + d_i\right), \quad \text{where} \quad d_i = \Phi^{-1}\left(1-\alpha_i/2\right)\sqrt{\frac{p_i\left(1-p_i\right)}{N}},$$

Or, equivalently:

$$p_i \pm \Phi^{-1}\left(1-\alpha_i/2\right)\sqrt{\frac{p_i\left(1-p_i\right)}{N}}. \qquad (2)$$

Above,

N is the number of trials (the sample size),
α_i is the *significance level* for this interval,
d_i is the half-width of the interval, and
Φ is the cumulative distribution function (CDF) for the standard normal distribution; Φ^{-1} is the inverse of the standard normal CDF[2].

The above interval is called a two-sided $100(1-\alpha_i)\%$ confidence interval for π_i. The proper interpretation of the interval $(p_i-d_i,\ p_i+d_i)$ is as follows: in hypothetical repetitions of the experiment, if equation (2) were always used to construct a confidence interval, the intervals so constructed would, on average, enclose the true value of π_i about $100(1-\alpha_i)\%$ of the time[3].

[2] The quantity $\Phi^{-1}(1-\alpha/2)$ is the upper $100(\alpha/2)$th percentile of the standard normal distribution; it is sometimes written as $z_{\alpha/2}$.

[3] The interval formed by (2) is the standard one used for binomial proportions when the sample size is moderately large and the true proportion is not too close to 0 or 1. It is based on a normal approximation. Other intervals could be formed based on the true underlying binomial distribution if necessary. See, for example, the Clopper-Pearson interval described in, e.g., Hollander and Wolfe (1999).

The quantity $100(1-\alpha_i)\%$ is called the *confidence level*, or the *coverage probability*, for the interval. For a given α-level, the width of the interval (or its half-width, d_i) may be considered a measure of the precision with which the estimate p_i approximates the true proportion π_i.

Example

Q: Say that a sample of N = 100 people are tested. 27 of them are found to be in group BB, while 73 are in other groups. What is the 95% confidence interval (α = 0.05) for the proportion π_{BB}?

A: The estimated proportion is p_{BB} = 0.27. Using equation (2), the confidence interval is (0.183, 0.357). The half-width d_{BB} is 0.087. So at the 0.05 significance level, the precision of the estimated proportion is plus/minus about 9 percentage points.

The Multiple Comparisons Problem

The standard confidence interval (2) could be applied to all 16 estimated proportions to provide upper and lower bounds. This would provide some measure of precision for each estimate, and would thus constitute a major improvement over only reporting the point estimates. Furthermore, equation (2) involves N, so it is possible to use it to try to determine the required sample size.

When dealing with multiple proportions, the above comments are subject to two major complications:

1. If each of the 16 intervals were done with the same significance level α, then the coverage probability of all 16 intervals *as a group* will be much lower than $1-\alpha$. For example, if all intervals are constructed to have 95% coverage probability, then the probability that all 16 intervals will actually cover the true π is much smaller than 95%. This is called the *multiple comparisons problem* or the problem of *simultaneous testing*.
2. When we are constructing 16 intervals, their coverage probabilities and half-widths will all depend on the common sample size N. So it is not clear how to choose a single value for N that will be small, but will still give good properties for all of the intervals.

Because of these two problems, the determination of sample size is not completely straightforward. A number of methods have been proposed to solve this problem; two of them, representing different approaches, will be discussed in the next section. To help understand these methods, two concepts must be reviewed.

The concept of groupwise significance level. The significance level for an individual confidence interval has been represented by α_i. The groupwise significance level will be denoted α. For a set of k confidence intervals, the interpretations of α_i and α are

$$\alpha_i = P(\text{interval i fails to contain its true proportion } \pi_i)$$
$$\alpha = P(\text{at least one of the k intervals fails to contain its true proportion}).$$

When doing many tests (k = 16 in our case), it is important to control the groupwise level α. For example, the intervals may be designed so that, on hypothetical repetitions of the experiment, there is only a 5% chance that even one of the intervals misses its true proportion. To achieve this, it will be necessary to control each α_i to a level considerably below 5%.

The Bonferroni correction. The Bonferroni inequality is used to relate the significance level of the individual tests to the groupwise significance level. The essential result is that, for k tests:

$$\alpha \leq \sum_{i=1}^{k} \alpha_i \qquad (3)$$

That is, the sum of the significance levels of the individual intervals provides an upper bound on the groupwise significance level. For example, if five intervals are constructed, each at level $\alpha_i = 0.01$, then the groupwise significance level α will definitely be smaller than 0.05. Based on inequality (3), it is customary to set the individual significance level for each of k tests to α/k.

Two Approaches to Calculating Sample Size

Below, two methods from the literature are used to explore the question of sample size for the upcoming experiment. Both methods are reasonably easy to implement, but nontrivial to understand. As such, the required formulas and description are presented, without further discussion of how or why the methods work. It is hoped that the concepts explained above should render the methods understandable.

Option 1: Tortora Method

This method is from Tortora (1978), and is also discussed in Bromaghin (1993). It is useful in the following situations:

- When it is desired to control the α_i for each interval, as well as the groupwise significance level α.
- When it is possible to specify good guesses for the vector of population proportions, $\boldsymbol{\pi}$.
- When it is advantageous to use the standard formula given in (2) to form the confidence intervals.

Inputs:

- A desired groupwise significance level, α.
- A set of required confidence level half-widths, d_i (i = 1, 2, …, 16).
  - Default may be to assign the same precision d to all d_i.
- A set of individual significance levels, α_i.
  - Default may be to assign equal levels $\alpha_i = \alpha/16$.
- A set of estimates for the true proportions π_i.

- Note: the calculated sample size depends on the choice of π, so it is important to have a good guess of the proportions. If no good guess is available, several options can be tested, and the largest sample size can be taken as a worst case.

Calculation of sample size:

The principle of this method is simple. The sample size required is calculated based on each proportion individually, and then the largest estimate is taken as the final result.

Let n_i be the sample size estimate determined from the information for the i^{th} proportion.

1. Calculate each n_i, i = 1, 2, …, 16 as follows:

$$n_i = \text{ceil}\left[\frac{\pi_i(1-\pi_i)}{d_i^2}\left(\Phi^{-1}(1-\alpha_i/2)\right)^2\right] \quad (4)$$

   where ceil() is the ceiling (round up) function.
2. Set the final sample size to N = max(n_i).

Note that equation (4) is simply equation (2) solved for N.

An example calculation:

First, let the desired groupwise error rate be α. One problem with using equation (4) is that separate values of π_i, d_i, and α_i must each be specified for all 16 proportions. For present purposes, the problem will be simplified by forcing certain relationships among the π_i's and d_i's.

Specification of true proportions, π_i:

- Assume the four correct diagnoses (AA, BB, CC, DD) have the same probability, call it π_{ii}. Correct diagnoses are expected to dominate, so that $4\pi_{ii}$ should be large (say, 0.8-0.9).
- Assume the remaining 12 incorrect diagnoses also have equal probability, π_{ij}. Once π_{ii} is specified, π_{ij} is determined, since $4\pi_{ii} + 12\pi_{ij} = 1$.

Specification of desired precisions, d_i:

- Set the half-width of the intervals for AA, BB, CC, DD to a common value d_{ii}, and then set the other 12 half-widths to $2d_{ii}$. This implies that the proportions should be measured with twice the precision for the important cases AA, BB, CC, DD.

Specification of individual significance levels:

- Assume all intervals should be calculated with the same significance level $\alpha^* = \alpha/16$.

Making the above assumptions means that the full sets of π_i, d_i, and α_i values needed for equation (4) can be specified through three values: α, π_{ii}, and d_{ii}.

Example case: $\alpha = 0.05$, $\pi_{ii} = 0.2$, $d_{ii} = 0.05$

- The value of $\pi_{ii} = 0.2$ implies that the total probability of successful classification is $\pi_{AA} + \pi_{BB} + \pi_{CC} + \pi_{DD} = 0.8$, and then that the 12 misclassification probabilities are all $\pi_{ij} = 0.0167$. The intervals for the correct classifications will look like $p_{ii} \pm 0.05$, and for the misclassifications, $p_{ij} \pm 0.10$. Each of the intervals has significance level $\alpha^* = 0.003125$.

These parameter choices and the calculated n_i from equation (4) are summarized below.

Table 2. Example calculation from the Tortora method.

i	Diag-noses	Inputs			Output
		π_i	d_i	α_i	n_i
1	AA	0.2	0.05	0.003125	559
2	BB	0.2	0.05	0.003125	559
3	CC	0.2	0.05	0.003125	559
4	DD	0.2	0.05	0.003125	559
5	AB	0.0167	0.10	0.003125	15
6	AC	0.0167	0.10	0.003125	15
7	AD	0.0167	0.10	0.003125	15
8	BA	0.0167	0.10	0.003125	15
9	BC	0.0167	0.10	0.003125	15
10	BD	0.0167	0.10	0.003125	15
11	CA	0.0167	0.10	0.003125	15
12	CB	0.0167	0.10	0.003125	15
13	CD	0.0167	0.10	0.003125	15
14	DA	0.0167	0.10	0.003125	15
15	DB	0.0167	0.10	0.003125	15
16	DC	0.0167	0.10	0.003125	15

The largest ni calculated was 559, so the result of this calculation is to recommend a sample size of N = 559 for the experiment.

The large sample size required in this example is due to the relatively narrow confidence intervals required for AA, BB, CC, and DD. Experience has shown that the interval half-widths d_i dominate the sample size calculation. If a very narrow confidence interval is demanded, the required sample size will grow quite rapidly.

Figure 1, below, shows the results of the Tortora method calculated exactly as above, but for different combinations of α and d_{ii}. This figure can be used to study the precision-sample size tradeoffs in more detail. Note again that these results depend on the particular set of π_i's chosen, so that additional calculations using (4) are required for other scenarios.

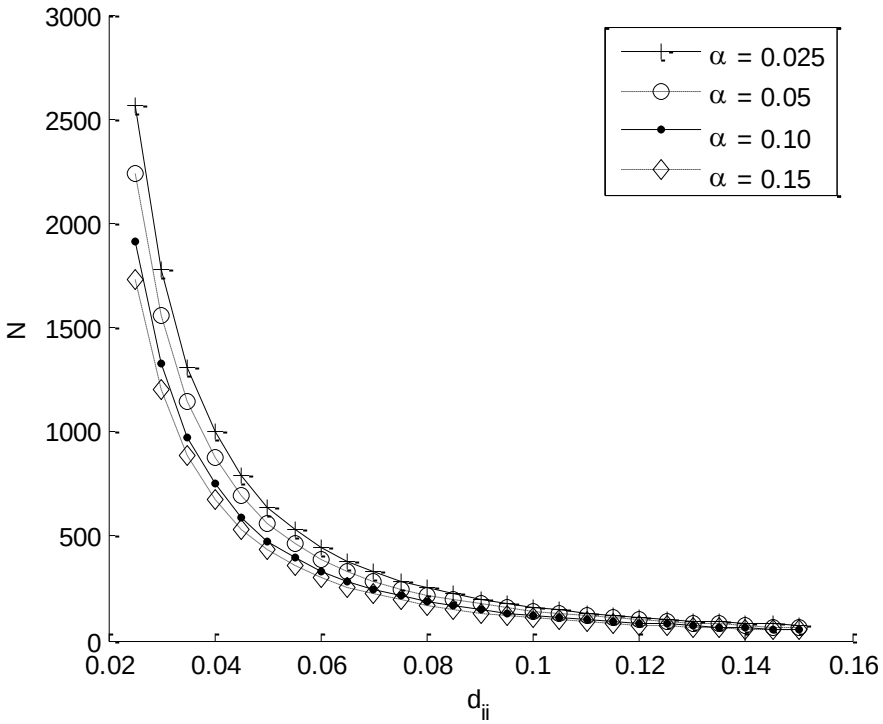

Figure 1. Results of the Tortora method. Sample size vs. interval half-width for the conditions described in the text.

Option 2: Thompson Method

The Tortora method just described is particularly useful if it is important to specify the minimum coverage of each individual interval, *and* specify a certain groupwise coverage rate. The Thompson method (Thompson, 1987; Bromaghin, 1993), to be described next, has three characteristics that differentiate it from the previous method:

1. It is based on intervals that *all have the same width*, and are of the form $p_i \pm d$. So only a single half-width d must be chosen, and subsequent reporting of the proportions only need mention that "all proportions are estimated plus or minus d."
2. Because the intervals are all equally wide, the significance level of each interval is not controlled to pre-specified values. *Only the groupwise significance level is controlled*. So only a single overall level α needs to be specified.
3. It does not require specification of π_i's. The method calculates a sample size that will be satisfactory even for the least favourable combination of probabilities.

These characteristics make the calculation of sample size somewhat less cluttered than the Tortora method.

Inputs:

- The desired groupwise significance level, α.

· The desired interval half-width, d.

Calculation of sample size:

The sample size is determined from equation (5):

$$N = \text{ceil}\left(\max_m \left[\left(\Phi^{-1}\left(1 - \alpha/2m\right)\right)^2 \frac{m-1}{m^2 d^2} \right] \right), \qquad (5)$$

Where m is an integer. The expression in the square brackets is to be maximized over all integers m, but in practice the maximum will occur at a small value of m, after which the result will continuously decrease. So it is only necessary to start at m=1 and try increasing m until the result begins to decrease.

Thompson (1987) also provides a small table that makes evaluation of N particularly easy for the given values of α, for any d. A portion of the table is reproduced below.

Table 3. A quick method of calculation for the Thompson method at particular α values. Divide the appropriate entry in the d^2N column by the desired value of d to find N.

α	$\mathbf{d^2N}$
0.20	0.74739
0.10	1.00635
0.05	1.27359
0.025	1.55963

An example calculation:

Equation (5) can be used to produce a figure equivalent to Figure 1, but using the Thompson method. The results are shown in Figure 2. Note that unlike the Tortora approach, Figure 2 applies to any particular true set of probabilities that may exist.

Figure 2 agrees with Figure 1 in the general shape and placement of the curves. The two methods of calculating sample size largely corroborate one another.

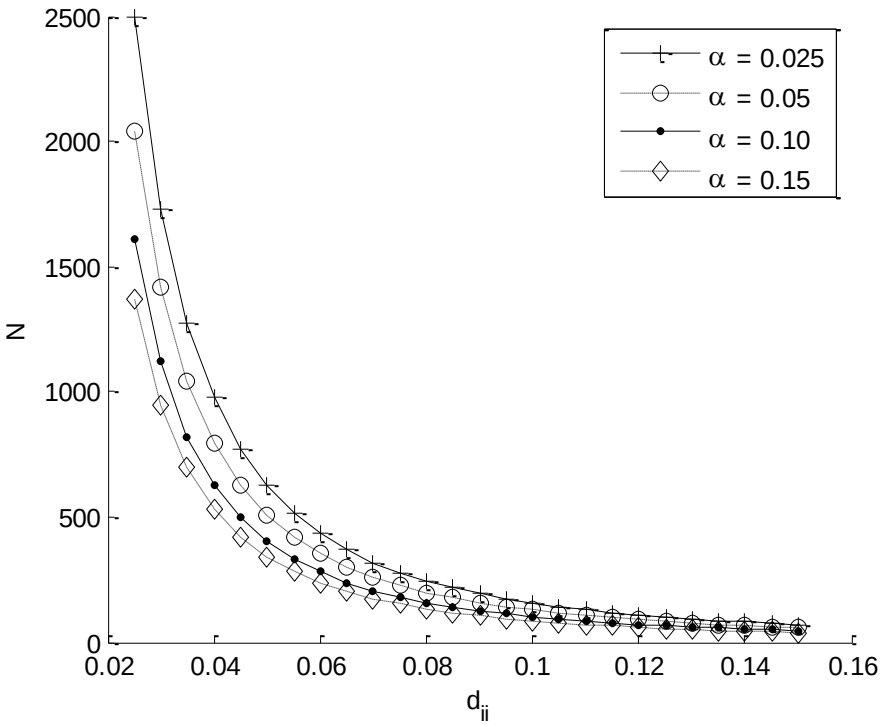

Figure 2. Results of the Thompson method. Sample size vs. interval half-width for different groupwise significance levels.

Summary and Conclusions

Despite the apparent complexity of the preceding analyses, the conclusions of this study are relatively straightforward.

What sample size is recommended?

This question cannot be answered conclusively without more understanding of what constitutes a "big" and "small" sample size in the practical context of the experiment. The two methods of estimating sample size agree reasonably well, so that either method (and either Figure 1 or Figure 2) can be used to evaluate the trade-offs between sample size and precision. The following general comments can be made:

- A half-width of d = 0.05 is probably a good goal for any proportion that needs to be estimated with high precision.
- If the goal is d = 0.05 with a good groupwise error rate, then a sample size of approximately 500 is needed.
- If a sample size in the hundreds is considered unacceptable, the only viable solution is to considerably increase the acceptable groupwise significance level (or to ignore the multiple comparisons problem completely). If this approach is taken, the investigators should be aware that one or more of the confidence intervals constructed are likely to miss their true values.

After choosing N and getting the data, how should data analysis be done?

The most important message is to report a confidence interval for any proportions that are calculated and reported. Equation (2) can be used as a general equation for constructing an interval for a proportion. Reporting an interval estimate rather than just a point estimate is a vast improvement.

Beyond just using intervals, some consideration of the multiple comparisons issue would be another improvement. This can be done in one of two ways:

a) By using intervals calculated with equation (2), but using a Bonferroni-corrected significance level (α/k).

b) By using a fixed-width interval as discussed in the Thompson method, and reporting only the appropriate groupwise significance level. Table 3, for example, could be used to determine the matching α, d, and N values.

Appendix: Ignoring the Multiple Comparisons Problem for the Purposes of Choosing Sample Size

Another way to think of the study is to assume that each patient has been pre-screened for a single pathology. The tests for each patient then become a single yes/no decision, so that only one proportion (the proportion of "yes" results) needs to be estimated.

In this situation one could consider the set of patients in each pathology group as essentially a separate study, and just construct each confidence interval at the usual α = 0.05 level—without correction for multiple comparisons[4]. If intervals are constructed using equation (2), then equation (4) can be used to determine the approximate sample size needed in each pathology group.

The results of such a sample size calculation are shown in Figure 3, for different combinations of π and d, with a fixed α value of 0.05. It is clear from the figure that the required sample size has a strong dependence on both the true proportion and the desired half-width. The dashed vertical line in the figure is drawn at π = 0.90, a plausible value from historical data. At this value of π, a sample size of 35 is needed to obtain a confidence interval with a half-width of d = 0.1.

[4] Note that viewing the problem in this way does not eliminate the multiple comparisons problem; it just eliminates it from discussion. If 95% confidence intervals are made for the "yes" proportion for four different pathologies, the chance that *all* intervals cover their true values could be as low as 80%. This is not necessarily a problem, and should not influence the interpretation of each individual interval; but it is worth being aware of this effect whenever considering several statistical tests or intervals at once.

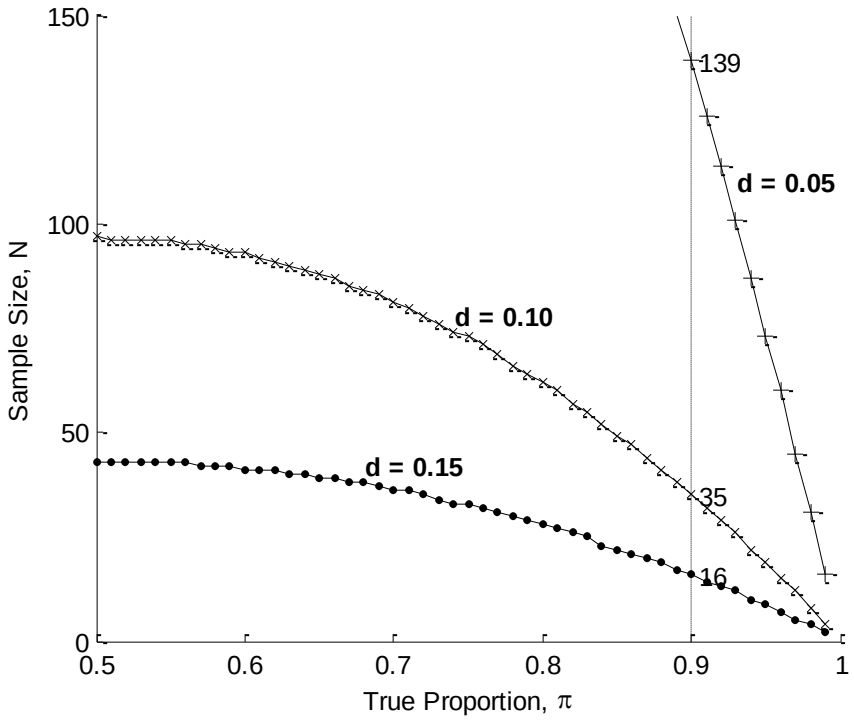

Figure 3. Sample sizes required for estimation of a single proportion, as a function of π, the true proportion, and d, the half-width of a 95% confidence interval. A vertical line, with calculated N values shown, is drawn at the plausible value $\pi = 0.9$.

References

Bromaghin, J. F. (1993), "Sample Size Determination for Interval Estimation of Multinomial Probabilities," The American Statistician, 47(3), 203-206.

Hollander, M. and Wolfe, D. A. (1999), *Nonparametric Statistical Methods* (2nd ed.), Wiley.

Tortora, R. D. (1978), "A Note on Sample Size Estimation for Multinomial Populations," *The American Statistician*, 32(3), 100-102

Thompson, S. K. (1987), "Sample Size for Estimating Multinomial Proportions," *The American Statistician*, 41(1), 42-46.