

Sample Size Calculations for Test and Evaluation

Alyson G. Wilson

Cowboy Programming Resources, Inc., El Paso, Texas

Sample size calculations can help testers and evaluators decide how many samples they need to collect to perform an experiment. These methods provide an explicit way of trading off risk against cost. This paper presents the ideas needed to perform sample size calculations, formulas for the most common Gaussian and binomial data cases, and examples of their use.

The primary purpose of statistical experimentation is to collect information to support decision making. Every time an experiment is performed, the experimenter must decide how many samples he or she can afford to collect and analyze, given limited resources. Further, the experimenter must decide whether available resources are sufficient to provide the information needed to support the decision. There is a constant trade-off between how few samples are needed in order to collect adequate information and how many samples can be collected. These trade-offs can be assessed using statistical methods of sample size calculation.

To provide concrete examples for calculating sample size, consider the following scenario. An air defense system needs to be tested and evaluated. There are two questions of interest:

- (1) What is the probability that the system detects eligible targets?
- (2) What is the downrange error in the system's estimate of a target's launch point?

After initial testing, improvements will be made in the system, and additional testing will then be performed to make sure that system performance has not been degraded. How can appropriate sample sizes be planned for all these tests?

Decision-making guidelines

There are four steps in establishing decision-making guidelines before an experiment is conducted: stating the alternative decisions, defining the acceptable risks for selecting the wrong alternative, establishing an objective criterion for selecting between the alternative decisions, and computing the requisite sample size (Diamond 1981).

Stating Alternatives

The first steps in stating alternative decisions are to identify the quantity of interest and to decide what sort of data will be collected to create information about that quantity. This sounds like a simple problem, but that is not always the case. Suppose one wants to assess the effectiveness of a new training program. What *measurable* quantity captures the idea of effectiveness? Could you give a multiple choice pre-test and post-test and compare scores? Could you measure how quickly a trainee performs a task before and after training?

TYPES OF DATA

The type of data collected helps determine the appropriate statistical methods for sample size calculations. This paper considers two of the most common types of data. The first type of data is Bernoulli/binomial data. This data arises from two-choice situations: yes or no, heads or tails, detect or not detect. The data is often presented as a proportion. For example, in the scenario mentioned in the introduction, question 1 can be addressed by examining what proportion of the time eligible targets were detected.

The second type of data is Gaussian or normal data. The distribution of Gaussian data looks like a "bell curve." Gaussian data often arises when a continuous measurement is made: e.g., How tall are these soldiers? What is the tensile strength of this type of steel? Question 2 from the scenario could be planned assuming Gaussian data.

ESTIMATION VERSUS TESTING

After the type of data to be collected is determined, the type of answer required from the data must be chosen. The two most common goals of an experiment are esti-

mation and testing. In an estimation problem, one is interested in using experimental data to estimate some unknown quantity and provide an indication of how good the estimate is. For example, one might want to collect data to estimate how often a system can detect a target or how long it takes a part to fail. In an estimation problem, one is not interested in making a comparison—only in providing summary numbers.

In a testing problem, one is interested in comparing a quantity of interest against some criterion. For example, one might be interested in whether the probability of detection for a particular system is at least 0.7, or in whether system improvements increase detection range as compared to the old system.

In a testing situation, one makes two additional decisions. The first is whether a *one-sample* or a *two-sample* procedure is required. One-sample procedures require one set of data and comparison against a fixed criterion. For example, is the probability of detection for this system greater than 0.7? Two-sample procedures compare two sets of data. Is the new system's probability of detection greater than the old system's probability of detection?

The second decision is whether a *one-sided* or *two-sided* test is needed. A one-sided test is used when one is interested in whether the quantity of interest is greater than (or less than) some criterion. For example, does this system detect farther out than the old system? Is the downrange error less than 0.3 km? A two-sided test is used when one is interested in whether the quantity of interest is different (not equal) to some criterion. Is the detection range 4 km, or does it vary either larger or smaller?

SUMMARY OF DECISIONS REQUIRED

The decisions required to state alternatives can be summarized as:

- What problem am I addressing, and what sort of data will I collect to address it?
- Is the data Bernoulli/binomial, Gaussian, or some other type?
- Am I interested in estimation or testing?
- If I am interested in testing, do I have a one-sample or a two-sample problem?
- If I am interested in testing, do I have a one-sided or a two-sided test?

Defining acceptable risks

Two types of errors can be made in the test and evaluation setting. The first is deciding that a system meets its operational requirements when it does not. This results in a bad system going to the field. The second is deciding that system does not meet requirements when it actually does. This results in spending a lot of money and not fielding a worthwhile system. The trade-off between these risks drives many of the decisions in sample size

calculation. Defining acceptable risks requires a different thought process for estimation and testing.

ACCEPTABLE RISKS FOR ESTIMATION

In an estimation problem, the results are usually expressed using a confidence interval. A confidence interval is a range of values calculated from the observed data that tries to capture the quantity of interest. The basic idea of a confidence interval is that if an experiment were performed a large number of times and a confidence interval were formed each time, $100(1 - \alpha)$ percent of the time, the confidence interval would contain the quantity that we are trying to estimate. (Notice that we cannot say anything about whether this particular confidence interval contains the parameter—only that confidence intervals *in general* contain the parameter the appropriate proportion of the time.)

The first quantity that we have to specify is the α mentioned above. This value is often known as the level of significance of the confidence interval, and it is typically set at 0.01, 0.05, or 0.1. Choosing $\alpha = 0.05$ corresponds to choosing a 95% confidence interval. The second quantity to be chosen is the precision, d , of the estimate. For a confidence interval, d corresponds to half of the width of the confidence interval. For example, is it sufficient to estimate the detection range to within plus or minus $d = 50$ meters? To get more precision requires more samples.

ACCEPTABLE RISKS FOR TESTING

Defining acceptable risks for testing requires more steps. The first step is to write down formally the hypotheses being examined. The null hypothesis (H_0) assumes there is no essential difference between the quantity of interest and the criterion. For example, H_0 : The proportion of detections of eligible targets is 0.7. The alternative hypothesis (H_A) assumes that there is a difference. The type of difference was determined in the "Estimation versus Testing" paragraph above by deciding whether a one-sided or two-sided test was needed. For example, for a one-sided test, H_A : The proportion of detections of eligible targets is less than 0.7.

There are two types of incorrect decision that can be made based upon these hypotheses. Defining acceptable risk implies choosing a probability that these mistakes will be made. The first type of error is called *Type 1* or α error. A Type 1 error occurs when one rejects H_0 when H_0 is true. For example, one concludes that the proportion of detections of eligible targets is less than 0.7 when it is actually 0.7. Traditionally, Type 1 error is set at 0.01, 0.05, or 0.1, although it can be set at any value between 0 and 1.

Type 2 or β error occurs when one accepts H_0 when H_A is true. For example, one concludes that the probabil-

ity of detection is 0.7 when it is actually smaller. *Power* is 1 - the probability of type 2 error, or the probability of rejecting H_0 when it is false. *Table 1* summarizes the types of error. Effect size (ES), listed in the truth column, is discussed in the next paragraph.

TABLE 1. Type 1 and Type 2 Error

TRUTH	DECISION	
	No Improvement	Improvement
No Improvement	Correct Decision Probability: $1 - \alpha$	Type 1 Error Probability: α
Improvement of Effect Size	Type 2 Error Probability: β	Correct Decision Probability: $1 - \beta$ (Power)

Establishing objective criteria

When calculating sample size, it is necessary to specify the minimum acceptable degree of change, or the *effect size* (ES), that one wants to detect. Sample size calculations are dependent upon specifying an exact value for the ES, not simply "any change." For example, suppose that in a baseline test, the probability of detection was 0.7. For a system enhancement, we want to be sure that performance has not degraded. What is the minimum value by which the new data must differ from the baseline data to be considered a degradation? In order to calculate a sample size for comparison, one must specify a specific ES to detect; for example, any change of greater than 5% should be detected as a degradation. If one looks at sample size as a function of effect size, Type 1 error, and Type 2 error, the smaller the effect size one wants to detect, the larger the required sample size.

Computing sample size

Many of the formulas for computing sample size use percentile points of the normal distribution. In the notation of the formulas, z_p is the $100(1 - p)$ percentile point of the normal distribution. Common points are listed in *Table 2*.

TABLE 2: Percentiles of the normal distribution

p	100(1-p) percentile
0.010	2.326
0.025	1.960
0.050	1.645
0.100	1.282
0.150	1.036
0.200	0.842
0.250	0.675
0.300	0.524

The following sections discuss sample size formulas for specific types of data. Within each group, there is a

discussion of an estimation problem, a one-sample test(s), and a two-sample test(s).

BERNOULLI/BINOMIAL DATA

Bernoulli/binomial data arises when the response variable has two levels: yes or no, detect or not detect. The quantity of interest is the proportion of "success," p , where one of the responses is designated as success. In the estimation problem, one is trying to get a numerical estimate of p ; in the one-sample testing problem, one is comparing p against a fixed criterion; in the two-sample testing problem, one is comparing the proportions arising from two data sets.

Estimation of Proportion. There are two quantities that must be specified: α , the level of significance of the confidence interval, and d , the width of the confidence interval. In addition, one must specify a guess for the proportion of success: call this guess p_0 . The formula for sample size, n , is given by:

$$n = \left(\frac{z_{\alpha/2}}{d} \right)^2 p_0 (1 - p_0) \quad (1)$$

If it is not possible to specify a guess for p_0 , or if one wishes to use a conservative estimate, then notice that the maximum value for $p_0(1 - p_0)$ occurs when $p_0 = 0.5$. In this case, the formula simplifies to:

$$n = \frac{z_{\alpha/2}^2}{4d^2} \quad (2)$$

Consider question 1 from the scenario in the introduction. Suppose that we would like to form a 90% confidence interval for the proportion of detections of eligible targets that has a total width of 10% (i.e., plus or minus 5%). Suppose also that we do not have any baseline data that would allow us to estimate p_0 . Using the second formula, with $z_{\alpha/2} = 1.645$ and $d = 0.05$, we find a sample size of $n = 271$. If we could estimate $p_0 = 0.7$, then using the first formula, we find a sample size of $n = 227$. By more precisely specifying p_0 , one can plan a test with fewer samples.

One-sample Test of Proportion. In order to calculate the sample size for a one-sample testing problem, one needs to specify both the null and alternative hypotheses. For Bernoulli/binomial data, the null hypothesis has the form $H_0: p = p_0$ where p_0 is the base case criterion value. The alternative hypothesis depends on whether a one-sided or two-sided test is required. Choose p_1 so that the effect size is $|p_1 - p_0|$.

One must specify the Type 1 error rate (α) and the Type 2 error rate (β). There are two formulas that one

can use to calculate the sample size (n), and both should give similar results. Arcsin is inverse sine, with its argument in radians. The formulas are given for a one-sided test. To perform a two-sided test, substitute $z_{\alpha/2}$ for z_{α} , but not $z_{\beta/2}$ for z_{β} .

$$n = \left(\frac{z_{\alpha} \sqrt{p_0(1-p_0)} + z_{\beta} \sqrt{p_1(1-p_1)}}{p_1 - p_0} \right)^2 \quad (3)$$

Alternately,

$$n = \left[\frac{z_{\alpha} + z_{\beta}}{2(\arcsin \sqrt{p_1} - \arcsin \sqrt{p_0})} \right]^2 \quad (4)$$

Consider again question 1 from the scenario in the introduction. Suppose that we would like to compare the proportion of detection against a criterion of 0.7 using a one-sided test, with $H_0: p = 0.7$ and $H_A: p < 0.7$. We would like a Type 1 error of 10%, a power of 80%, and an effect size of 5%. To use the formulas, we have $z_{\alpha} = 1.282$, $z_{\beta} = 0.842$, $p_0 = 0.7$, and $p_1 = 0.65$. Using the first formula, we have $n = 391$; using the second formula, we have $n = 395$. For planning purposes, these sample sizes are essentially the same.

Testing the of Equality of Two Proportions. A common problem in two-sample testing is the case where a previous experiment has been performed with a sample size of n_0 and we want to calculate how many samples we should use for a subsequent test (n_1) to achieve a certain level of Type 1 and Type 2 error. To use the formula below, the null hypothesis has the form $H_0: p = p_0$ and the effect size is $|p_1 - p_0|$, but p_0 is the proportion from the initial experiment. As before, to perform a two-sided test, substitute $z_{\alpha/2}$ for z_{α} , but not $z_{\beta/2}$ for z_{β} .

The formula is given as:

$$n_1 = \frac{mn_0}{2n_0 - m} \quad (5)$$

where

$$m = 2 \left(\frac{z_{\alpha} + z_{\beta}}{2(\arcsin \sqrt{p_1} - \arcsin \sqrt{p_0})} \right)^2 \quad (6)$$

If the denominator of the first equation is zero or negative, then the problem is not solvable for the given specifications. One must either increase n_0 (usually impossible) or change the desired power, Type 1 error, or effect size to decrease m .

Consider question 1 from the scenario in the introduction. Suppose that we ran an initial test with $n_0 = 400$ and $p_0 = 0.7$. Our hypotheses are one-sided, with

$H_0: p_0 = p_1$ and $H_A: p_0 > p_1$. In other words, we want to be sure that there is no degradation with the system improvements. We would like a Type 1 error of 10%, a power of 80%, and an effect size of 5%. To use the formulas, we have $z_{\alpha} = 1.282$, $z_{\beta} = 0.842$, and $p_1 = 0.65$. Using the formula, we find $n_1 = 31,640$.

Clearly something strange is happening here. We used a somewhat larger sample size than was required for the one-sample test, but the sample size required for the subsequent experiment has exploded. The explanation for this is that when one performs a two-sample test, one needs to have approximately twice as many observations in each group as are required for the single group in the one-sample test. The additional samples are required to account for the variability in the estimate obtained from the baseline case. If a follow-up test is planned, more samples need to be collected for the baseline case to ensure that a good comparison can be made with a reasonable number of samples.

Consider question 1 again, but suppose that we ran an initial test with $n_0 = 750$ and the same values discussed above. Using the formula, we find $n_1 = 837$.

GAUSSIAN DATA

Gaussian data arises when a continuous measurement is made that results in data that fall roughly in the shape of a bell curve. The quantity of interest in our discussion is the average or mean, μ , of the data, which roughly corresponds to the highest point on a histogram of the data. The mean measures the center of the data. Although the variance of the data will be unknown in some of our problems, this article will not address variance estimation.

In the estimation problem, one is trying to get a numerical estimate of μ ; in the one-sample testing problem, one is comparing μ against a fixed criterion; in the two-sample testing problem, one is comparing the means arising from two data sets.

Estimation of Mean. The first case of interest focuses on the estimation of the mean of Gaussian data. There are two quantities that must be specified: α , the level of significance of the confidence interval, and d , the width of the confidence interval.

Estimation of Mean with Variance Known. In the case where the variance, σ^2 , of the underlying Gaussian distribution for the data is known, it can be used directly in the formula for sample size (n).

$$n = \left(\frac{z_{\alpha/2} \sigma}{d} \right)^2 \quad (7)$$

Consider question 2 from the scenario in the introduction. Suppose that we would like to form a 90% confidence interval that had a total width of 3 km, and suppose also that we knew (probably from

prior experience) that the standard deviation of the data was 8 km. Then we have $z_{\alpha/2} = 1.645$, $\sigma = 8$, and $d = 1.5$, which implies that $n = 77$.

Estimation of Mean with Variance Unknown. If the variance of the underlying Gaussian distribution is unknown, then the most straightforward way to estimate a sample size is to specify the precision d as a multiple of the standard deviation, e.g., $d = k\sigma$. Then the formula in the above paragraph simplifies to:

$$n = \left(\frac{z_{\alpha/2}}{k} \right)^2 \quad (8)$$

Suppose that we would like to form a 90% confidence interval, but that we could not estimate the standard deviation of the data's underlying distribution. Suppose also, however, that we would be happy with a confidence interval with a half-width of 0.5 standard deviations. Then $k = 0.5$, $z_{\alpha/2} = 1.645$, and $n = 11$. This sample size is smaller than that in the above example of "estimation of mean with variance known" because there $k = d/\sigma = 0.1875$, and more precision was required.

One-Sample Test of Mean. A one-sample test of the mean involves comparing the mean against some fixed criterion. Again, there are two cases: variance known and variance unknown.

One-Sample Test of Mean with Variance Known. For Gaussian data, the null hypothesis has the form $H_0: \mu = \mu_0$, where μ_0 is the base case criterion value. The alternative hypothesis depends on whether a one-sided or two-sided test is required. Choose μ_1 so that the effect size is $|\mu_1 - \mu_0|$.

One must specify the Type 1 error rate (α) and the Type 2 error rate (β). Since the variance, σ^2 , of the underlying Gaussian distribution for the data is assumed to be known, it can be used directly in the formula for sample size (n). As usual, to perform a two-sided test, substitute $z_{\alpha/2}$ for z_α , but not $z_{\beta/2}$ for z_β .

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{|\mu_1 - \mu_0|} \right]^2 + 1 \quad (9)$$

Using the formula above, suppose that we have a standard deviation σ of 8, an effect size of 1, and desire Type 1 and Type 2 error rates of 10%. To perform a one-sided test would require a sample size of $n = 422$.

One-Sample Test of Mean with Variance Unknown. The case where the variance of the normal distribution is unknown is considerably harder. There are a variety of approaches (see, for example, Desu and Raghavarao 1990, p. 9-11). Two of the simplest approaches are presented here, and both require iterative calculations.

Suppose that σ_u^2 is a known upper bound for the unknown variance σ^2 , and choose μ_1 so that the effect size is $|\mu_1 - \mu_0|$. Then it can be shown that the appropriate sample size for a one-sided test is the smallest positive integer satisfying the following equation (Desu and Raghavarao 1990), where $t_{m,p}$ is the 100(1-p) percentile point of the t distribution with m degrees of freedom:

$$n = \left[\frac{\sigma_u(t_{n-1,\alpha} + t_{n-1,\beta})}{\mu_1 - \mu_0} \right]^2 \quad (10)$$

Of course, it is probably unrealistic to assume that a good upper bound can be chosen for an unknown variance. If the upper bound is chosen to be too large, then the sample size is unnecessarily large.

Another procedure is proposed in Diamond (1981, p. 55). First specify the effect size as a multiple of the standard deviation so that $|\mu_1 - \mu_0| = k\sigma$. To get a starting value, substitute into equation number (9) above, ignoring the final +1 term, as shown below.

$$n_{(0)} = \left(\frac{z_\alpha + z_\beta}{k} \right)^2 \quad (11)$$

Take this value of $n_{(0)}$ and substitute it into the following equation, which is a simplification of the equation given in the previous method.

$$n_{(1)} = \left(\frac{t_{n-1,\alpha} + t_{n-1,\beta}}{k} \right)^2 \quad (12)$$

Continue the iterations by substituting $n_{(1)}$ into the second equation to get $n_{(2)}$, etc., until the result does not change.

Testing the Equality of Two Means. A common problem in two-sample testing is the case where a previous experiment has been performed with a sample size of n_0 and we want to calculate how many samples we should use for a subsequent test (n_1) to achieve a certain level of Type 1 and Type 2 error. For normal data, assume that the null hypothesis has the form $H_0: \mu = \mu_0$ and the effect size is $|\mu_1 - \mu_0|$, but μ_0 is the mean from the initial experiment.

Testing the Equality of Two Means when the Variances are Known. If the variances for both the first and the second set of data are known, then the sample size can be calculated by substituting and solving for n_1 . As before, to perform a two-sided test, substitute $z_{\alpha/2}$ for z_α , but not $z_{\beta/2}$ for z_β .

$$\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} = \left(\frac{\mu_1 - \mu_0}{z_\alpha + z_\beta} \right)^2 \quad (13)$$

As an example, suppose that a preliminary experiment had been done, with $n_0 = 75$, $\sigma_0 = 3$, $\mu_0 = 2$. Setting Type 1 and Type 2 error rates at 10%, considering an effect size of 1, assuming $\sigma_1 = \sigma_0$, and considering a one-sided test gives $n_1 = 280$.

Testing the Equality of Two Means when the Variances are Unknown but Assumed Equal. Again, the variance unknown case requires an iterative solution. If the effect size can be specified as $d = k\sigma$, then the new sample size n_1 can be calculated by solving for the smallest positive integer satisfying:

$$\frac{n_0 n_1}{n_0 + n_1} = \left(\frac{t_{n_0+n_1-2, \alpha} + t_{n_0+n_1-2, \beta}}{k} \right)^2 \quad (14)$$

As before, to perform a two-sided test, substitute $z_{\alpha/2}$ for z_α , but not $z_{\beta/2}$ for z_β .

Testing the Equality of Two Means with the Variances Unknown but Not Assumed Equal. This problem, known as the Behrens-Fisher problem, falls beyond the scope of this paper. See Desu and Raghavarao (1990) for a discussion of possible approaches.

Conclusions

The following general principles apply to sample size calculations and are adapted from Kraemer and Theimann (1987). These principles provide intuition about the ways in which changes in error rates, effect sizes, and sample sizes are interrelated.

- The smaller the probability of Type 1 (α) error desired, the larger the necessary sample size (n). More samples are needed to achieve a lower error probability.
- The smaller the probability of Type 2 (β) error desired, the larger the necessary sample size (n). More samples are needed to achieve a lower error probability.
- For a fixed sample size and effect size, lowering the Type 1 error probability increases the Type 2 error probability, and vice versa.
- For a fixed effect size and Type 1 error probability, the smaller the sample size, the smaller the power (i.e., the larger the Type 2 error probability).
- Two-tailed tests require larger sample sizes than one-tailed tests. A two-tailed test examines two directions at once, and consequently requires more samples.
- The smaller the effect size, the larger the required sample size. Detection of small changes requires more data than detection of large changes.
- If the proposed sample size is smaller than 20, one must be willing to either tolerate high error probabilities or be working in an area where the effect size is quite large.

This paper addresses common situations where sample sizes can be calculated analytically. For more complicated situations, there are a variety of software

packages available. Recent reviews can be found in Thomas (1997) and Thomas and Krebs (1997). □

References

- Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences, Revised Edition*. New York: Academic Press.
- Desu, M. and Raghavarao, D. 1990. *Sample Size Methodology*. San Diego, Calif.: Academic Press.
- Diamond, W. 1981. *Practical Experiment Designs for Engineers and Scientists*. Belmont, Calif.: Wadsworth.
- Fleiss, J. 1973. *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.
- Kraemer, H. and Theimann, S. 1987. *How Many Subjects?* Newbury Park, Calif.: Sage Publications.
- Thomas, L. November 30, 1997. "A Comprehensive List of Power Analysis Software for Microcomputers." *Statistical Power Analysis Software*. <http://www.interchg.ubc.ca/cacb/power>.
- Thomas, L. and Krebs, C. J. 1997. "A Review of Statistical Power Analysis Software." *Bulletin of the Ecological Society of America*, 78(2), 128-139.

Alyson Gabbard Wilson is a statistician with Cowboy Programming Resources, Inc., in El Paso, Texas, specializing in the operational evaluation of air defense artillery systems. Her Ph.D. is from the Institute of Statistics and Decision Sciences, Duke University. Dr. Wilson is a member of the American Statistical Association, the International Society for Bayesian Analysis, and the International Test and Evaluation Association. Dr. Wilson also has work and consulting experience in the pharmaceutical industry and in biomedical research.

SPROLES – REFERENCES (CONTINUED FROM PAGE 30)

- Macquarie 1985. *The Macquarie Dictionary*. Sydney: Macquarie Library Pty. Ltd.
- Oxford Advanced Learners Dictionary of Current English*. 1989. Oxford: Oxford University Press.
- Morse, P. M. and Kimball, G. E. 1950. *Methods of Operations Research*. Washington, DC: U.S. Navy OEG Report.
- Oxford English Dictionary, Second Edition*. 1989. Vol. V, Oxford: Clarendon Press.
- Pinker, A., Samuel, A. H., and Batcher, R. December 1995. "On Measures of Effectiveness." *Phalanx*, The Journal of the Military Operations Research Society, 8-12.
- Roche, J. G. and Watts, B. D. June 1991. "Choosing Analytic Measures." *The Journal of Strategic Studies*, Vol. 14, 165-209.
- Simon, H. A. 1984. *The Sciences of the Artificial*, Second Edition, Third Reprint. Cambridge, Massachusetts: The MIT Press.