Given a significance test, How large a sample size is large enough?

Mark E. Chittenden Jr. a

a College of William and Mary, Virginia Institute of Marine Science, Gloucester Point


To cite this article: Mark E. Chittenden Jr. (2002): Given a significance test, How large a sample size is large enough?, Fisheries, 27:8, 25-29

To link to this article: http://dx.doi.org/10.1577/1548-8446(2002)027<0025:GASTHL>2.0.CO;2
Given a significance test, how large a sample size is large enough?

"The sample size $n$ is too small" is an often-devastating criticism widely expressed in evaluating research. Researchers are often not aware, however, that, for reasons described herein, whether $n$ be 2 or 2,000, rejection of the null hypothesis ($H_0$) by a significance test immediately establishes that $n$ is acceptably adequate to support the conclusion(s) that must then be drawn. Those conclusions amount to accepting an alternative hypothesis given the protocols followed in evaluating $H_0$. Traditional fears of small $n$, however, really are warranted when $H_0$ is not rejected.

Introduction

That the sample size $n$ is too small is a criticism widely expressed in evaluating research. That criticism leads to the basic question addressed herein: "How large a $n$ is large enough?" That question is one of the most important things addressed in planning and evaluating research. Researchers usually begin to encounter the principles of this topic in texts for courses in introductory statistics like Sokal and Rohlf (1995) and Steel et al. (1997). More advanced treatment occurs in texts for courses on sampling, experimental design, and analysis of variance like Thompson (1992) and Zar (1996).

Despite all the available information, researchers are often unaware of basic principles that must be understood to judge what is or is not adequate $n$. This article describes principles that explain why, given a significance test, rejection of the null hypothesis immediately establishes that $n$ is adequate to support the conclusions that must be drawn. In doing so, I describe the principles and protocols of the null and alternative hypotheses, Type I and Type II errors, and significance tests, and how significance levels are related to the probability of committing Type I and Type II errors. I present a minimum of mathematics, because I hope to reach not researchers with a strong background in statistics and sample size estimation but, rather, those who have a weaker background and might be deterred by math.

It is my hope that, if researchers understand the principles described, unwarranted criticism that a $n$ is too small to support the conclusion(s) drawn will cease where the null hypothesis is rejected by a significance test. That would force criticisms to focus on more appropriate issues and better advance our science.

The Null and Alternative Hypotheses and Their Protocols

One of the first steps in the decision-making process implied by a significance test is to pose one or more hypotheses to be tested. That step is based on preliminary information, and some concern over the population from which the sample is drawn. Taking an imaginary population as an example, say the lengths of the fish in a pond, let us make an initial hypothesis that the mean length in the population, $\mu_0$, is 50 mm. That initial hypothesis is referred to as the null hypothesis, $H_0$, and the respective subscripts refer to a hypothesized mean and a null hypothesis. We will draw a sample of size $n$ of fish from the population, measure each fish, and calculate a sample mean length to make some decision about $H_0$. In doing so, we actually draw a sample from a sampling frame, not the population. The sampling frame and the population are different, though quite parallel, concepts. We ignore the distinction between them for simplicity.

Quite often $H_0$ is about the last thing we believe. Rather, it is a "strawman" that is compared against the sample mean and evaluated by a significance test. The primary objective of the test is actually to reject $H_0$. That perspective occurs because often we work backwards in a significance test, e.g., we can not know with absolute certainty that $H_0$ is correct unless we can somehow examine the entire population. We can conclude, however, that $H_0$ is almost certainly not correct if we have sufficient evidence, the criterion being that $H_0$ is rejected by the test. If we do not have sufficient evidence, $H_0$ is not rejected. In the latter case, there are two possible conclusions: 1) $H_0$ may be correct, or 2) more refined study is needed to reject $H_0$. We come back to these latter cases at the end of this section.

Associated with $H_0$ are one or two alternative hypotheses, $H_{alt}$ or $H_1$ and $H_2$, where the subscripts 1 and 2 refer to alternatives to $H_0$. One of the alternatives must be accepted by default if $H_0$ is rejected by the test. An alternative

Mark E. Chittenden, Jr.

Chittenden is a professor of marine science at the College of William and Mary, Virginia Institute of Marine Science, Gloucester Point. He can be reached at markc@vims.edu.
hypothsis, however, is often the one we really believe.

Generally, there are two alternative hypotheses. In our example, they would be: 1) the mean length in the population is >50 mm, and 2) the mean is <50. The significance test is referred to as a two-tailed test when two alternative hypotheses are of interest. In our example, we would accept the first alternative if the sample mean was >50, the second if it was <50.

If only one alternative hypothesis has interest or meaning, we have a one-tailed test. Examples include: 1) when water or fish samples are examined to see if they contain levels of some hazardous contaminant, like PCBs, that exceed an accepted tolerance level, that is taken as the \( H_0 \) to be tested, and 2) when drinking water samples are examined to see if they contain levels of some beneficial additive, like chlorine, that fall below an accepted target level, that is taken as \( H_0 \). The single \( H_{alt} \) in these cases would be that the population mean level exceeds the tolerance level (1) or falls below the target level (2).

The previous description primarily addresses the case where the significance test rejects \( H_0 \). It is important to understand also the implications when \( H_0 \) is not rejected. In that case, we often see the conclusion worded to the effect that “we accept” or reject \( H_0 \). Symbols in parentheses indicate probabilities of making the decision at given \( n \).

**Type I and Type II Errors**

Only four outcomes are possible when a significance test is made, two correct and two erroneous (Figure 1): 1) to “accept” a true \( H_0 \), a correct conclusion, 2) to reject a true \( H_0 \), an erroneous conclusion referred to as a Type I or \( \alpha \) error, 3) to “accept” a false \( H_0 \), an erroneous conclusion referred to as a Type II or \( \beta \) error, and 4) to reject a false \( H_0 \), a correct conclusion.

Once a decision is made, only one type of error is possible as the columns in Figure 1 show. If the decision is to reject \( H_0 \), only a Type I error is possible. If the decision is to “accept,” only a Type II error is possible.

The probabilities of making Type I and Type II errors are denoted as \( \alpha \) and \( \beta \) respectively (Figure 1). Subtracting from one, the probability of “accepting” a true \( H_0 \) is \( 1 - \alpha \), and the probability of rejecting a false \( H_0 \), referred to as the power of the test, is \( 1 - \beta \).

Although both \( \alpha \) and \( \beta \) provide avenues for hypothesis testing (Freund and Wilson 1993), \( \alpha \) generally has been used on the premise that Type I error is of more concern than Type II. We describe below how the probability is quantified and used in making significance tests. Many introductory statistics texts treat \( \beta \), especially Dixon and Massey (1969). Peterman (1990) treats \( \beta \) in the context of power and fisheries examples, but see also Hoenig and Heisey (2001) who warn against a posteriori use of power calculations.

### Sampling Distribution of the Mean and Principles of Significance Tests

A significance test may be viewed as a set of rules that we follow to evaluate an \( H_0 \) and decide whether to reject it or not. The rules are based on the properties of: 1) the distribution of the sample mean, and 2) the distribution of observations from which the sample means are obtained. These properties illustrate how we decide to “accept” or reject \( H_0 \) and how Type I error and significance levels are related.

We start with a simple observation, \( y_i \) or \( x_i \), an individual, usually numerical recording of a piece of information. Each observation comes from an overall set of observations, termed a population. In our example, the observations would be the individual lengths of the fish in the pond, and the population would be made up of all the lengths.

The values of all the observations in the population may be arranged as a frequency distribution (Figure 2A). This population of observations is referred to as a parent population or distribution to distinguish it from the distribution of sample means.
derived from it. The population of observations is characterized by several fixed parameters: a population mean, \( \mu \), a population variance, \( \sigma^2 \), and a population standard deviation, \( \sigma \). The population mean, \( \mu \), which describes central tendency, is usually of most interest.

We draw a sample to estimate and evaluate parameters of the parent population. In doing so, we use sample statistics to estimate the parameters: the sample mean, \( \bar{y} \) or \( \bar{x} \), to estimate \( \mu \); the sample variance, \( s^2 \), to estimate \( \sigma^2 \); and the sample standard deviation, \( s \), to estimate \( \sigma \). In trying to estimate the fixed parameters, we have the problem that sample statistics vary, depending on the particular sample of observations we draw from the population. The properties of the sampling distribution of the mean help us get around that problem as follows.

If we select from the parent population all possible samples (APS) of size \( n \) and calculate the mean, \( \bar{y} \), for each sample, we can arrange the values of those sample means as a frequency distribution (Figure 2B) like we did for the observations. This second frequency distribution is the sampling distribution of the mean, and it nestles within the distribution of observations it is derived from. The two distributions are identical when \( n = 1 \).

The sampling distribution of the mean has the important properties that: 1) the distribution of the sample means becomes more and more like a normal distribution as \( n \) increases, the Central Limit Theorem, 2) the distribution of the sample means is normal if the parent population is normal, regardless of \( n \), 3) the mean of the sample means of APS of size \( n \), \( \mu_{\bar{y}} \), is equal to the mean of the parent population, \( \mu \), 4) the variance of the sampling distribution of the mean, \( \sigma_{\bar{y}}^2 \), is \( \sigma^2/n \), and 5) the standard deviation of the sampling distribution of the mean, \( \sigma_{\bar{y}} \), the standard error of the mean, is \( \sqrt{\sigma^2/n} \). Taken together, 1 and 2 let us use properties of the normal distribution in testing significance, and 3–5 describe the parameters actually used.

Parameters of the sampling distribution of the mean can be estimated by sample statistics similar to how we estimated parent population parameters: \( \mu_{\bar{y}} \) by \( \bar{y} \), the mean of the sample means; \( \sigma_{\bar{y}}^2 \) by \( s^2 \)

---

**Figure 2.** Frequency distributions of A) observations \( y_i \) from the parent population and B) sample means \( \bar{y} \) of all possible samples of size \( n \) from the parent population. Population means of these respective distributions, \( \mu \) and \( \mu_{\bar{y}} \), equal 50 to satisfy the text example.

---

**Figure 3.** Diagrammatic representation of the relative frequency (rf) distribution of the sample means \( \bar{y} \) of all possible samples of size \( n \). \( \mu_{\bar{y}} \) is a hypothesized value for \( \mu_{\bar{y}} = \mu \). Upper and lower critical regions (shaded) indicate \( \bar{y} \)-values leading to rejection (shaded area) or "acceptance" (unshaded area) of \( H_0 \) at \( \alpha = 0.05 \). The two shaded areas, one in each tail, indicate principles of a two-tailed significance test. Both shaded areas are pooled into one tail for a one-tailed test.
are substituted for tie constants 1.96 and 2.576, depending on the unknown \( \mu \) and \( \sigma \) respectively. The statistics used to delineate the upper and lower critical regions (Figure 3).

The former intervals give the two 99% confidence intervals for \( \mu \), referred to as an “acceptance” region. The conclusion would not be so clear, however, for \( \bar{y} \)'s of 45 and 55. Significance tests are most useful when the conclusion is not clear, because they provide an objective basis for “accepting” or rejecting \( \mu_0 \).

The objective basis on which to “accept” or reject \( \mu_0 \) is simple. Certain regions, referred to as critical or rejection regions, are delineated on the sampling distribution of the mean (shaded areas in Figure 3). Sample means that fall within a critical region result in rejection. Means that fall outside a critical region result in “acceptance,” so that area (the unshaded area) is often referred to as an “acceptance” region.

The shaded area(s) represents the significance level chosen for making the test, e.g., the significance level would be 5% if the shaded area made up 5% of the area under the curve. In a two-tailed test, half the 5%—2.5%—would be at the end of each tail of the distribution (Figure 3). In a one-tailed test, the entire 5% would be located in either the upper or the lower tail (not both) depending on \( \alpha \).

Several significance levels are accepted as providing sufficient evidence for making a decision, especially the 1% and 5% levels. The significance level represents \( \alpha \), the probability of making a Type I error, so there is a 1% and a 5% probability of making a Type I error with these respective significance levels. Because the probability of “accepting” \( \mu_0 \) is 1—\( \alpha \) (Figure 1), there is a 99% and 95% probability of “accepting” a true \( \mu \), with these significance levels. The probability \( \alpha \) can also be viewed as the percentage of decisions that will result in a Type I error, e.g., only 1% and 5% of the decisions will be to reject a true \( \mu_0 \) at \( \alpha = 0.01 \) and \( \alpha = 0.05 \), respectively.

Delineation of the critical regions is simple. In a normal distribution, the intervals \( \mu_0 \pm 1.96\sigma \) and \( \mu_0 \pm 2.576\sigma \) include 95% and 99% of the observations, respectively. The intervals \( \mu_0 \pm 1.96\sigma_\bar{y} \) and \( \mu_0 \pm 2.576\sigma_\bar{y} \) include 95% and 99% of the sample means, respectively. The latter intervals give the two \( \bar{y} \)-values used to delineate the upper and lower critical regions (Figure 3).

The constants 1.96 and 2.576 are related to relative cumulative frequency units of a normal distribution with \( \mu = 0 \) and \( \sigma = 1 \). Being relative, they hold for all normal distributions, regardless of their means and standard deviations. These units are denoted as \( z \) or \( u \) where \( z = u = (y - \mu) / \sigma \) and \( z = u = (y - \mu_\bar{y}) / \sigma_\bar{y} \) for the parent population and the distribution of sample means, respectively. The statistics \( s \) and \( s_\bar{y} \) are used to estimate the usually unknown \( \sigma \) and \( \sigma_\bar{y} \). In that case, \( t \)-values from the \( t \)-distribution are substituted for the constants 1.96 and 2.576, depending on the degrees of freedom of \( s \) or \( s_\bar{y} \).

**Discussion**

The description above provides the basis for clear answers to the question, “Given a significance test, how large a \( n \) is large enough to support a conclusion to reject \( \mu_0 \)?”

The significance test provides an accepted objective basis, a set of rules, on which to decide if there is sufficient evidence to reject \( \mu_0 \). That decision basically depends on the magnitude of \( \bar{y} \) in relation to \( \mu_0 \). Once a decision is made, only one type of error is possible. If the decision is to reject, only a Type I error can be made, rejection of a true \( \mu_0 \). Regardless of \( n \), be it 2 or 2000, a defined but acceptably low and unavoidable fraction of the decisions to reject will commit a Type I error. That fraction is the significance level. The 1% and 5% significance levels generally used correspond to 1% and 5% probabilities, respectively, of rejecting a true \( \mu_0 \). As such, they permit only acceptably low probabilities of rejecting a true \( \mu_0 \), though a 0% level does not exist in sampling. Therefore, though it may be surprising, as long as a significance test is made, any \( n \), be it 2 or 2000, is large enough to acceptably support a decision to reject \( \mu_0 \). Having rejected \( \mu_0 \), an \( H_{alt} \) must be accepted by default. Given these principles, criticisms that \( n \) is too small are not valid and should be avoided when \( \mu_0 \) is rejected by a significance test. The lower limit on \( n \) in that case is that it must provide at least one degree of freedom, so that a significance test can be made. It is obvious that the test assumptions must hold.

If the decision is to “accept” \( \mu_0 \), only a Type II error can be made, “accepting” a false \( \mu_0 \). In that case, small \( n \) is dangerous. That is so because at a given significance level, the power of the test, the probability of rejecting a false \( \mu_0 \) (1 — \( \beta \)), is proportional to \( n \)—it is low with small \( n \) but increases with large \( n \). The problem of small \( n \) is lessened if \( \mu_0 \) is far from the true mean \( \mu \), it is magnified if \( \mu_0 \) is close to \( \mu \). All in all, small \( n \) may lead to “accepting” almost any \( \mu_0 \). That case should be regarded with caution, because it is the specific case where traditional fears of small \( n \) really are warranted.

Nothing said herein should be taken as a recommendation to plan to use small \( n \), because small \( n \) generally results in wide confidence intervals, and a high probability of Type II error when \( \mu_0 \) is “accepted.” Often it is not possible, however, to achieve a larger intended \( n \). In those cases, results are useful when a significance test rejects \( \mu_0 \).

It is wise to use standard methods to estimate required \( n \) before collecting data. Generally simple, they ask the researcher to specify beforehand: 1) the maximum error that he/she can tolerate in estimating some measure of central tendency, 2) an acceptable probability that that error may be exceeded, and 3) an estimate of the variance. These requirements are generally not difficult in practice. The first two simply reflect the researcher’s tolerance for error, and many simple methods exist to estimate the variance. Further description of sample-size estimation is beyond the scope of this article, but it is well-explained in many sampling and experimental design texts.

**Acknowledgements**

Financial support was provided by the College of William and Mary, Virginia Institute of Marine Science and by a Wallop/Breaux Program Grant from the U.S. Fish and Wildlife Service through the Virginia Marine Resources Commission, Project No. F-88-R3. This paper is Contribution No. 2468 from the College of William and Mary, Virginia Institute of Marine Science, Gloucester Point.
References


