2

Sample Size

The first question faced by a statistical consultant, and frequently the last, is, "How many subjects (animals, units) do I need?" This usually results in exploring the size of the treatment effects the researcher has in mind and the variability of the observational units. Researchers are usually less interested in questions of Type I error, Type II error, and one-sided versus two-sided alternatives. A key question is to settle the type of variable (endpoint) the consultee has in mind: Is it continuous, discrete, or something else? For continuous measurements the normal distribution is the default model, for distributions with binary outcomes, the binomial.

The ingredients in a sample size calculation, for one or two groups, are:

Type I Error (α)	Probability of rejecting the null hypothesis when it is true
Type II Error (β)	Probability of not rejecting the null hypothesis when it is false
Power = $1 - \beta$	Probability of rejecting the null hypothesis when it is false
σ_0^2 and σ_1^2	Variances under the null and alternative hypotheses (may be the same)
μ_0 and μ_1	Means under the null and alternative hypotheses
n_0 and n_1	Sample sizes in two groups (may be the same)

The choice of the alternative hypothesis is challenging. Researchers sometimes say that if they knew the value of the alternative hypothesis, they would not need to do the study. There is also debate about which is the null hypothesis and which is the

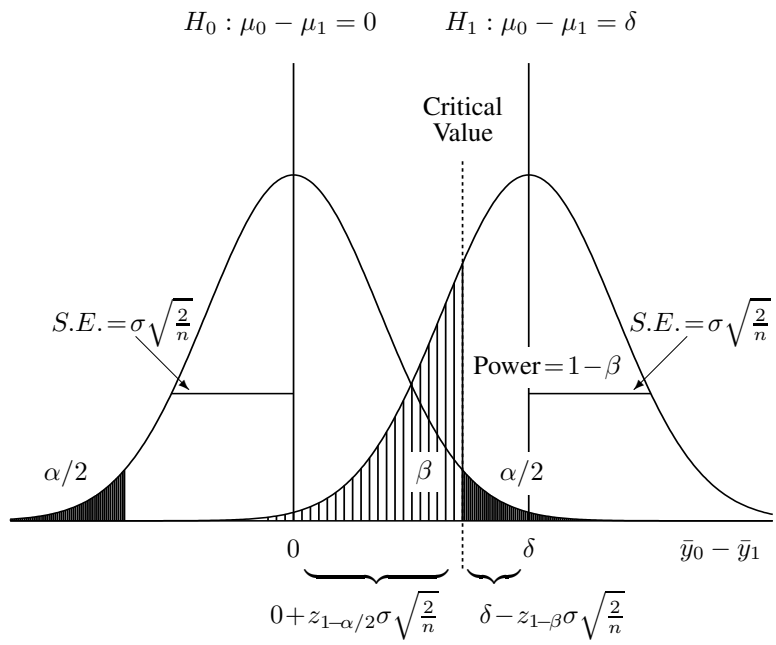

Fig. 2.1 Sampling model for two independent sample case. Two-sided alternative, equal variances under null and alternative hypotheses.

alternative hypothesis. The discussion can become quite philosophical, but there are practical implications as well. In environmental studies does one assume that a site is safe or hazardous as the null hypothesis? Millard (1987a) argues persuasively that the choice affects sample size calculations. This is a difficult issue. Fortunately, in most research settings the null hypothesis is reasonably assumed to be the hypothesis of no effect. There is a need to become familiar with the research area in order to be of more than marginal use to the investigator. In terms of the alternative hypothesis, it is salutary to read the comments of Wright (1999) in a completely different context, but very applicable to the researcher: "an alternative hypothesis ... must make sense of the data, do so with an essential simplicity, and shed light on other areas." This provides some challenging guidance to the selection of an alternative hypothesis.

The phrase, "Type I error," is used loosely in the statistical literature. It can refer to the error as such, or the probability of making a Type I error. It will usually be clear from the context which is meant.

Figure 2.1 summarizes graphically the ingredients in sample size calculations. The null hypothesis provides the basis for determining the rejection region, whether the test is one-sided or two-sided, and the probability of a Type I error (α)–the size of the test. The alternative hypothesis then defines the power and the Type II error (β). Notice that moving the curve associated with the alternative hypothesis to the

right (equivalent to increasing the distance between null and alternative hypotheses) increases the area of the curve over the rejection region and thus increases the power. The *critical value* defines the boundary between the rejection and nonrejection regions. This value must be the same under the null and alternative hypotheses. This then leads to the fundamental equation for the two-sample situation:

$$0 + z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}} = \delta - z_{1-\beta}\sigma\sqrt{\frac{2}{n}} \ . \tag{2.1}$$

If the variances, and sample sizes, are not equal, then the standard deviations in equation (2.1) are replaced by the values associated with the null and alternative hypotheses, and individual sample sizes are inserted as follows,

$$0 + z_{1-\alpha/2}\sigma_0\sqrt{\frac{1}{n_0} + \frac{1}{n_1}} = \delta - z_{1-\beta}\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \ . \tag{2.2}$$

This formulation is the most general and is the basis for virtually all two-sample sample size calculations. These formulae can also be used in one-sample situations by assuming that one of the samples has an infinite number of observations.

2.1 BEGIN WITH A BASIC FORMULA FOR SAMPLE SIZE–LEHR'S EQUATION

Introduction

Start with the basic sample size formula for two groups, with a two-sided alternative, normal distribution with homogeneous variances ($\sigma_0^2 = \sigma_1^2 = \sigma^2$) and equal sample sizes ($n_0 = n_1 = n$).

Rule of Thumb

The basic formula is

$$n = \frac{16}{\Delta^2} \ , \tag{2.3}$$

where

$$\Delta = \frac{\mu_0 - \mu_1}{\sigma} = \frac{\delta}{\sigma} \tag{2.4}$$

is the treatment difference to be detected in units of the standard deviation–the standardized difference.

In the one-sample case the numerator is 8 instead of 16. This situation occurs when a single sample is compared with a known population value.

Illustration

If the standardized difference, Δ, is expected to be 0.5, then $16/0.5^2 = 64$ subjects per treatment will be needed. If the study requires only one group, then a total of

Table 2.1 Numerator for Sample Size Formula, Equation (2.3); Two-Sided Alternative Hypothesis, Type I Error, $\alpha = 0.05$

Type II Error β	Power $1-\beta$ Power	Numerator for Sample Size Equation (2.3)	
		One Sample	Two Sample
0.50	0.50	4	8
0.20	0.80	8	16
0.10	0.90	11	21
0.05	0.95	13	26
0.025	0.975	16	31

32 subjects will be needed. The two-sample scenario will require 128 subjects, the one-sample scenario one-fourth of that number. This illustrates the rule that the two-sample scenario requires four times as many observations as the one-sample scenario. The reason is that in the two-sample situation two means have to be estimated, doubling the variance, and, additionally, requires two groups.

Basis of the Rule

The formula for the sample size required to compare two population means, μ_0 and μ_1, with common variance, σ^2, is

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\dfrac{\mu_0 - \mu_1}{\sigma}\right)^2} . \tag{2.5}$$

This equation is derived from equation (2.1). For $\alpha = 0.05$ and $\beta = 0.20$ the values of $z_{1-\alpha/2}$ and $z_{1-\beta}$ are 1.96 and 0.84, respectively; and $2(z_{1-\alpha/2} + z_{1-\beta})^2 = 15.68$, which can be rounded up to 16, producing the rule of thumb above.

Discussion and Extensions

This rule should be memorized. The replacement of 1.96 by 2 appears in Snedecor and Cochran (1980), the equation was suggested by Lehr (1992).

The two key ingredients are the difference to be detected, $\delta = \mu_0 - \mu_1$, and the inherent variability of the observations indicated by σ^2. The numerator can be calculated for other values of Type I and Type II error. Table 2.1 lists the values of the numerator for Type I error of 0.05 and different values of Type II error and power. A power of 0.90 or 0.95 is frequently used to evaluate new drugs in Phase III clinical trials (usually double blind comparisons of new drug with placebo or standard); see

Lakatos (1998). One advantage of a power of 0.95 is that it bases the inferences on confidence intervals.

The two most common sample size situations involve one or two samples. Since the numerator in the rule of thumb is 8 for the one-sample case, this illustrates that the two-sample situation requires four times as many observations as the one-sample case. This pattern is confirmed by the numerators for sample sizes in Table 2.1.

If the researcher does not know the variability and cannot be led to an estimate, the discussion of sample size will have to be addressed in terms of standardized units. A lack of knowledge about variability of the measurements indicates that substantial education is necessary before sample sizes can be calculated.

Equation (2.3) can be used to calculate detectable difference for a given sample size, n. Inverting this equation gives

$$\Delta = \frac{4}{\sqrt{n}}\,, \tag{2.6}$$

or

$$\mu_0 - \mu_1 = \frac{4\sigma}{\sqrt{n}}\,. \tag{2.7}$$

In words, the detectable *standardized* difference in the two-sample case is about 4 divided by the square root of the number of observations per sample. The detectable (non-standardized) difference is four standard deviations divided by the square root of the number of observations per sample. For the one-sample case the numerator 4 is replaced by 2, and the equation is interpreted as the detectable deviation from some parameter value μ. Figure 2.2 relates sample size to power and detectable differences for the case of Type I error of 0.05. This figure also can be used for estimating sample sizes in connection with correlation, as discussed in Rule 4.4 on page (71).

This rule of thumb, represented by equation (2.2), is very robust and useful for sample size calculations. Many sample size questions can be formulated so that this rule can be applied.

2.2 CALCULATING SAMPLE SIZE USING THE COEFFICIENT OF VARIATION

Introduction

Consider the following dialogue in a consulting session:

"What kind of treatment effect are you anticipating?"

"Oh, I'm looking for a 20% change in the mean."

"Mmm, and how much variability is there in your observations?"

"About 30%"

The dialogue indicates how researchers frequently think about relative treatment effects and variability. How to address this question? It turns out, fortuitously, that the question can be answered. The question gets reformulated slightly by considering

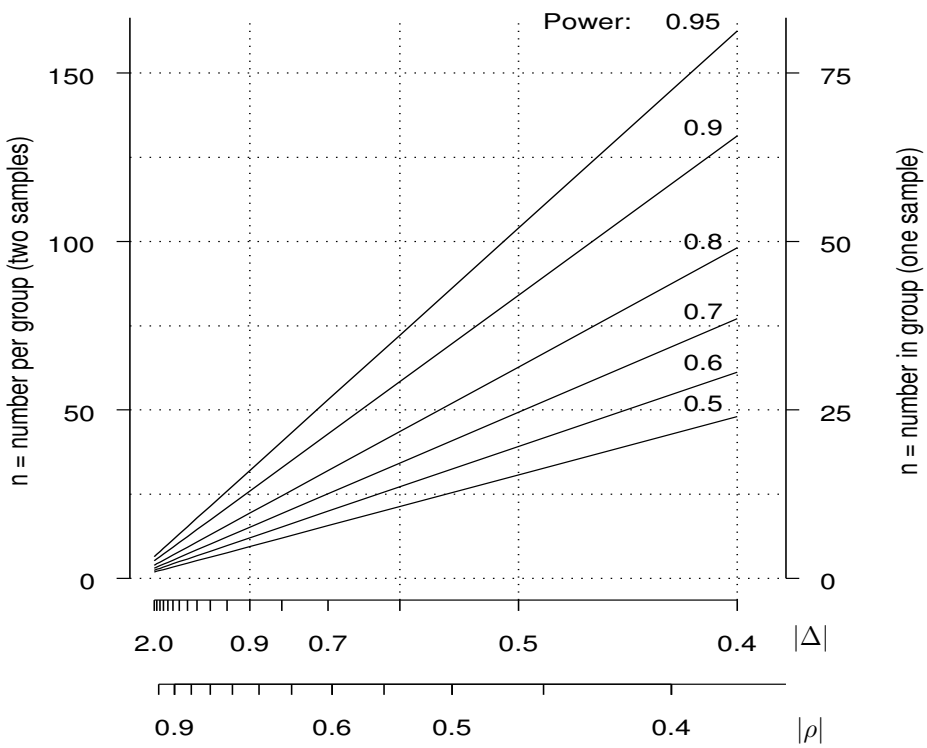

Fig. 2.2 Sample size for one sample and two samples. Use right-hand side for one-sample situation and correlation.

that the percentage change is linked specifically to the ratio of the means. That is,

$$\frac{\mu_0 - \mu_1}{\mu_0} = 1 - \frac{\mu_1}{\mu_0} \ . \tag{2.8}$$

. The question is then answered in terms of the ratio of the means.

Rule of Thumb

The sample size formula becomes:

$$n = \frac{16(CV)^2}{(\ln\mu_0 - \ln\mu_1)^2} \ , \tag{2.9}$$

where CV is the coefficient of variation ($CV = \sigma_0/\mu_0 = \sigma_1/\mu_1$).

Table 2.2 Sample Sizes for a Range of Coefficients of Variation and Percentage Change in Means: Two Sample Tests, Two-sided Alternatives with Type I Error 0.05 and Power 0.80 (Calculations Based on Equation 2.9)

Ratio of Means		**0.95**	**0.90**	**0.85**	**0.80**	**0.70**	**0.60**	**0.50**
% Change		**5**	**10**	**15**	**20**	**30**	**40**	**50**
	5	16	4	2	1	1	1	1
	10	61	15	7	4	2	1	1
Coefficient	**15**	137	33	14	8	3	2	1
of	**20**	244	58	25	14	6	3	2
Variation	**30**	548	130	55	29	12	6	3
in	**40**	974	231	97	52	21	10	6
Percent	**50**	>1000	361	152	81	32	16	9
	75	>1000	811	341	181	71	35	‘9
	100	>1000	>1000	606	322	126	62	34

Illustration

For the situation described in the consulting session ratio of the means is calculated to be $1 - 0.20 = 0.80$ and the sample size becomes

$$n = \frac{16(0.30)^2}{(\ln 0.80)^2} = 28.9 \approx 29.$$

The researcher will need to aim for about 29 subjects per group. If the treatment is to be compared with a standard, that is, only one group is needed, then the sample size required will be 15.

Basis of the Rule

Specification of a coefficient of variation implies that the standard deviation is proportional to the mean. To stabilize the variance a log transformation is used. In chapter 5 it is shown that the variance in the log scale is approximately equal to the coefficient of variation in the original scale. Also, to a first order approximation, the means in the original scale get transformed to the log of means in the log scale. The result then follows.

Discussion and Extensions

Table 2.2 lists the sample sizes based on equation (2.9) for values of CV ranging from 5% to 100% and values of PC ranging from 5% to 50%. These are ranges most likely to be encountered in practice.

Figure 2.3 presents an alternative way to estimating sample sizes for the situation where the specifications are made in terms of percentage change and coefficient of variation.

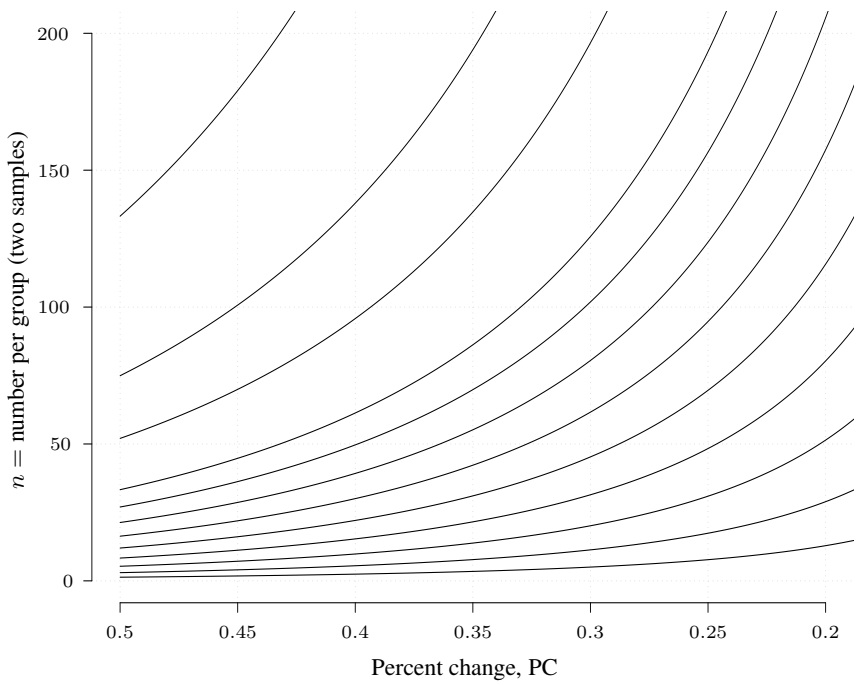

Fig. 2.3 Sample size for coefficient of variation and percent change. Conditions the same as for Table 2.2.

The percentage change in means can be defined in two ways using either μ_0 or μ_1 in the denominator. Suppose we define $\mu = (\mu_0 + \mu_1)/2$, that is, just the arithmetic average of the two means and define the quantity,

$$\text{PC} = \frac{\mu_0 - \mu_1}{\mu} \ . \tag{2.10}$$

Then the sample size is estimated remarkably accurately by

$$n = \frac{16 * CV^2}{PC^2} \ . \tag{2.11}$$

.

Sometimes the researcher will not have any idea about the variability. In biological systems a coefficient of variation of 35% is very common. A handy rule for sample size is then,

$$n = \frac{2}{PC^2} \ . \tag{2.12}$$

For example, under this scenario a 20% change in means will require about 50 subjects per sample. This is somewhat rough but a good first cut at the answers. Using equation 2.9 (after some algebra) produces $n = 49$

For additional discussion see van Belle and Martin (1993).

2.3 IGNORE THE FINITE POPULATION CORRECTION IN CALCULATING SAMPLE SIZE FOR A SURVEY

Introduction

Survey sampling questions are frequently addressed in terms of wanting to know a population proportion with a specified degree of precision. The sample size formula can be used with a power of 0.5 (which makes $z_{1-\beta} = 0$). The denominator for the two-sample situation then becomes 8, and 4 for the one sample situation (from Table 2.1).

Survey sampling typically deals with a finite population of size N with a corresponding reduction in the variability if sampling is without replacement. The reduction in the standard deviation is known as the finite population correction. Specifically, a sample of size n is taken without replacement from a population of size N, and the sample mean and its standard error are calculated. Then the standard error of the sample mean, $\bar{x}$ is

$$SE(\bar{x}) = \sqrt{\frac{N-n}{nN}}\sigma \ . \tag{2.13}$$

Rule of Thumb

The finite population correction can be ignored in initial discussions of survey sample size questions.

Illustration

A sample of 50 is taken without replacement from a population of 1000. Assume that the standard deviation of the population is 1. Then the standard error of the mean ignoring the finite population correction is 0.141. Including the finite population correction leads to a standard error of 0.138.

Basis of the Rule

The standard error with the finite population correction can be written as

$$SE(\bar{x}) = \frac{1}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\sigma \ . \tag{2.14}$$

So the finite population correction, $\sqrt{1 - \frac{n}{N}}$, is a number less than one, and the square root operation pulls it closer to one. If the sample is 10% of the population, the finite population correction is 0.95 or there is a 5% reduction in the standard deviation. This is not likely to be important and can be ignored in most preliminary reviews.

Discussion and Extensions

If the population is very large, as is frequently the case, the finite population correction can be ignored throughout the inference process. The formula also indicates that changes in orders of magnitude of the population will not greatly affect the precision of the estimate of the mean.

The rule indicates that initial sample size calculations can ignore the finite population correction, and the precision of the estimate of the population mean is proportional to $\sqrt{n}$ (for fixed standard deviation, σ).

2.4 THE RANGE OF THE OBSERVATIONS PROVIDES BOUNDS FOR THE STANDARD DEVIATION

Introduction

It is frequently helpful to be able to estimate quickly the variability in a small data set. This can be useful for discussions of sample size and also to get some idea about the distribution of data. This may also be of use in consulting sessions where the consultee can only give an estimate of the range of the observations. This approach is crude.

Rule of Thumb

The following inequalities relate the standard deviation and the range.

$$\frac{\text{Range}}{\sqrt{2(n-1)}} \leq s \leq \frac{n}{n-1}\frac{\text{Range}}{2} \ . \tag{2.15}$$

Illustration

Consider the following sample of eight observations: 44, 48, 52, 60, 61, 63, 66, 69. The range is $69 - 44 = 25$. On the basis of this value the standard deviation is bracketed by

$$2.6 \leq s \leq 14.3. \tag{2.16}$$

The actual value for the standard deviation is 8.9.

Basis of the Rule

This rule is based on the following two considerations. First, consider a sample of observations with range Range $= \mathrm{x}_{maximum} - \mathrm{x}_{minimum}$. The largest possible values for the standard deviation is when half the observations are placed at $\mathrm{x}_{minimum}$ and the other half at $\mathrm{x}_{maximum}$. The standard deviation in this situation is the upper bound in the equation. The smallest possible value for the standard deviation, given the same range, occurs when only one observation is at the minimum, one observation is at the maximum, and all the other observations are placed at the half-way point between the minimum and the maximum. For large n the the ratio $n/(n-1)$ is approximately 1 and can be ignored. So the standard deviation is less than Range$/2$.

Discussion and Extensions

The bounds on the standard deviation are pretty crude but it is surprising how often the rule will pick up gross errors such as confusing the standard error and standard deviation, confusing the variance and the standard deviation, or reporting the mean in one scale and the standard deviation in another scale.

Another estimate of the standard deviation is based on the range, assuming a normal distribution of the data, and sample size less than 15. For this situation

$$s \approx \frac{\text{Range}}{\sqrt{n}} \,. \tag{2.17}$$

For the example, Range$/\sqrt{n} = 25/\sqrt{8} = 8.8$ which is very close to the observed value of 8.9. This rule, first introduced by Mantel (1951) is based on his Table II. Another useful reference for a more general discussion is Gehan (1980).

2.5 DO NOT FORMULATE A STUDY SOLELY IN TERMS OF EFFECT SIZE

Introduction

The standardized difference, Δ, is sometimes called the *effect size*. As discussed in Rule 2.1, it scales the difference in population means by the standard deviation. Figure 2.1 is based on this standardized difference. While this represents a useful approach to sample size calculations, a caution is in order. The caution is serious enough that it merits a separate rule of thumb.

Rule of Thumb

Do not formulate objectives for a study solely in terms of effect size.

Illustration

Some social science journals insist that all hypotheses be formulated in terms of effect size. This is an unwarranted demand that places research in an unnecessary straightjacket.

Basis of the Rule

There are two reasons for caution in the use of effect size. First, the effect size is a function of at least three parameters, representing a substantial reduction of the parameter space. Second, the experimental design may preclude estimation of some of the parameters.

Discussion and Extensions

Insistence on effect size as a non-negotiable omnibus quantity (see Rule 1.10) forces the experimental design, or forces the researcher to get estimates of parameters from the literature. For example, if effect size is defined in terms of a subject-subject variance, then a design involving paired data will not be able to estimate that variance. So the researcher must go elsewhere to get the estimate, or change the design of the study to get it. This is unnecessarily restrictive.

The use of Δ in estimating sample size is not restrictive. For a given standard deviation the needed sample size may be too large and the researcher may want to look at alternative designs for possible smaller variances, including the use of covariates.

An omnibus quantity simplifies a a complicated parameter or variable space. This is desirable in order to get a handle on a research question. It does run the danger of making things too simple (Rule 1.9). One way to protect against this danger is to use effect size as one, of several, means of summarizing the data.

2.6 OVERLAPPING CONFIDENCE INTERVALS DO NOT IMPLY NONSIGNIFICANCE

Introduction

It is sometimes claimed that if two independent statistics have overlapping confidence intervals, then they are not significantly different. This is certainly true if there is substantial overlap. However, the overlap can be surprisingly large and the means still significantly different.

Rule of Thumb

Confidence intervals associated with statistics can overlap as much as 29% and the statistics can still be significantly different.

Illustration

Consider means of two independent samples. Suppose their values are 10 and 22 with equal standard errors of 4. The 95% confidence intervals for the two statistics, using the critical value of 1.96, are 2.2–17.8 and 14.2–29.8, displaying considerable overlap. However, the z-statistic comparing the two means has value

$$z = \frac{22 - 10}{\sqrt{4^2 + 4^2}} = 2.12.$$

Using the same criterion as applied to the confidence intervals this result is clearly significant. Analogously, the 95% confidence interval for the difference is $(22 - 10) \pm 1.96\sqrt{4^2 + 4^2}$, producing the interval 0.9–23.1. This interval does not straddle 0 and the conclusion is the same.

Basis of the Rule

Assume a two-sample situation with standard errors equal to σ_1 and σ_2 (this notation is chosen for convenience). For this discussion assume that a 95 % confidence level is of interest. Let the multiplier for the confidence interval for each mean be k to be chosen so that the mean difference just reaches statistical significance. The following equation must then hold:

$$\frac{k\sigma_1 + k\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} = 1.96. \tag{2.18}$$

Assuming that $\sigma_1 = \sigma_2$ leads to the result $k = 1.39$. That is, if the half-confidence interval is (1.39 $\times$ standard error) the means will be significantly different at the 5% level. Thus, the overlap can be $1 - 1.39/1.96 \sim 29\%$.

Discussion and Extensions

If the standard errors are not equal–due to heterogeneity of variances or unequal sample sizes–the overlap, which maintains a significant difference, decreases. Specifically, for a 95% confidence interval situation, if $r = \sigma_1/\sigma_2$ then

$$k = \sqrt{1 - \frac{r}{(r+1)^2}}\ 1.96. \tag{2.19}$$

This shows that the best that can be done is when $r = 1$. As r moves away from 1 (either way), the correction approaches 1, and k approaches 1.96.

The multiplier of 1.96 does not depend on the confidence level so the corrections apply regardless of the level. Another feature is that the correction involves a square root so is reasonably robust. For $r = 2$, corresponding to a difference in variances of 4, the overlap can still be 25%.

This rules indicates that it is insufficient to only consider non-overlapping confidence intervals to represent significant differences between the statistics involved. It may require a quick calculation to establish significance or non-significance. A

good rule of thumb is to assume that overlaps of 25% or less still suggest statistical significance. Payton et al. (2003) consider the probability of overlap under various alternative formulations of the problem.

2.7 SAMPLE SIZE CALCULATION FOR THE POISSON DISTRIBUTION

Introduction

The Poisson distribution is known as the law of small numbers, meaning that it deals with rare events. The term "rare" is undefined and needs to be considered in context. A rather elegant result for sample size calculations can be derived in the case of Poisson variables. It is based on the square root transformation of Poisson random variables.

Rule of Thumb

Suppose the means of samples from two Poisson populations are to be compared in a two-sample test. Let θ_0 and θ_1 be the means of the two populations. Then the required number of observations per sample is

$$n = \frac{4}{(\sqrt{\theta_0} - \sqrt{\theta_1})^2} \ . \tag{2.20}$$

Illustration

Suppose the hypothesized means are 30 and 36. Then the number of sampling units per group is required to be $4/(\sqrt{30} - \sqrt{36})^2 = 14.6 = 15$ observations per group.

Basis of the Rule

Let Y_i be Poisson with mean θ_i for $i = 0, 1$. Then it is known that $\sqrt{Y_i}$ is approximately normal ($\mu_i = \sqrt{\theta_i}, \sigma^2 = 0.25$). Using equation (2.3) the sample size formula for the Poisson case becomes equation (2.20).

Discussion and Extensions

The sample size formula can be rewritten as

$$n = \frac{4}{(\theta_0 + \theta_1)/2 - \sqrt{\theta_0 \theta_1}} \ . \tag{2.21}$$

The denominator is the difference between the arithmetic and the geometric means of the two Poisson distributions! The denominator is always positive since the arithmetic mean is larger than the geometric mean (Jensen's inequality).

Now suppose that the means θ_0 and θ_1 are means per unit time (or unit volume) and that the observations are observed for a period of time, *T*. Then Y_i is Poisson

with mean $\theta_i T$. Hence, the sample size required is

$$n = \frac{4}{T(\sqrt{\theta_0} - \sqrt{\theta_1})^2} \ . \tag{2.22}$$

This formula is worth contemplating. Increasing the observation period T, reduces the sample size proportionately, not as the square root! This is a basis for the observation that the precision of measurements of radioactive sources, which often follow a Poisson distribution, can be increased by increasing the duration of observation times.

Choose T so that the number per sample is 1. To achieve that effect choose T to be of length

$$T = \frac{4}{(\sqrt{\theta_0} - \sqrt{\theta_1})^2} \ . \tag{2.23}$$

This again is reasonable since the sum of independent Poisson variables is Poisson, that is, ΣY_i is Poisson $(T\theta_i)$ if each Y_i is Poisson (θ_i). This formulation will be used in Chapter 4, Rule 6.3 page (120), which discusses the number of events needed in the context of epidemiological studies.

The Poisson distribution can be considered the basis for a large number of discrete distributions such as the binomial, hypergeometric, and multinomial distributions. It is important to be familiar with some of these basic properties.

2.8 SAMPLE SIZE CALCULATION FOR POISSON DISTRIBUTION WITH BACKGROUND RATE

Introduction

The Poisson distribution is a common model for describing radioactive scenarios. Frequently there is background radiation over and above which a signal is to be detected. Another application is in epidemiology when a disease has a low background rate and a risk factor increases that rate. This is true for most disease situations. It turns out that the higher the background rate the larger the sample size needed to detect differences between two groups.

Rule of Thumb

Suppose that the background rate is θ^* and let θ_0 and θ_1 now be the additional rates over background. Then, Y_i is Poisson $(\theta^* + \theta_i)$. The rule of thumb sample size formula is

$$n = \frac{4}{(\sqrt{\theta^* + \theta_0} - \sqrt{\theta^* + \theta_1})^2} \ . \tag{2.24}$$

Illustration

Suppose the means of the two populations of radioactive counts are 1 and 2 with no background radiation. Then the sample size per group to detect this difference, using

equation (2.20), is $n = 24$. Now assume a background level of radiation of 1.5. Then the sample size per group, using equation (2.24), becomes 48. Thus the sample size has doubled with a background radiation halfway between the two means.

Basis of the Rule

The result follows directly by substituting the means $(\theta^* + \theta_0)$ and $(\theta^* + \theta_1)$ into equation (2.20).

Discussion and Extensions

It has been argued that older people survive until an important event in their lives has occurred and then die. For example, the number of deaths in the United States in the week before the year 2000 should be significantly lower than the number of deaths in the week following (or, perhaps better, the third week before the new year and the third week after the new year). How many additional deaths would need to be observed to be fairly certain of picking up some specified difference? This can be answered by equation (2.24). For this particular case assume that $\theta_0 = 0$ and $\theta_1 = \Delta\theta$. That is, $\Delta\theta$ is the increase in the number of deaths needed to have a power of 0.80 that it will be picked up, if it occurs. Assume also that the test is two-sided–there could be a decrease in the number of deaths. Suppose the average number of deaths per week in the United States (θ^*) is 50,000–a reasonable number. The sample size is $n = 1$. Some manipulation of equation (2.8) produces

$$\Delta\theta = 4\sqrt{\theta^*}\ . \tag{2.25}$$

For this situation an additional number of deaths equal to $4 \times \sqrt{50,000} = 894.4 \sim 895$ would be needed to be reasonably certain that the assertion had been validated. All these calculations assume that the weeks were pre-specified without looking at the data. Equation (2.25) is very useful in making a quick determination about increases in rates than can be detected for a given background rate. In the one-sample situation the multiplier 4 can be replaced by 2.

A second application uses equation (2.3) as follows. Suppose n^* is the sample size associated with the situation of a background rate of θ^*. Let $\bar{\theta} = (\theta_0 + \theta_1)/2$ be the arithmetic mean of the two parameters. Then using equation (2.3) for the sample size calculation (rather than the square root formulation) it can be shown that

$$n^* = n\left(1 + \frac{\theta^*}{\bar{\theta}}\right). \tag{2.26}$$

Thus, if the background rate is twice the average increase to be detected, then the sample size is doubled. This confirms the calculation of the illustration. The problem could also have been formulated by assuming the background rate increased by a factor R so that the rates are θ^* and $R\theta^*$. This idea will be explored in Chapter 4, Rule 6.3.

This rule of thumb is a good illustration of how a basic rule can be modified in a straightforward way to cover more general situations of surprising usefulness.

2.9 SAMPLE SIZE CALCULATION FOR THE BINOMIAL DISTRIBUTION

Introduction

The binomial distribution provides a useful model for independent Bernoulli trials. The sample size formula in equation (2.3) can be used for an approximation to the sample size question involving two independent binomial samples. Using the same labels for variables as in the Poisson case, let $Y_i,\ i = 0, 1$ be independent binomial random variables with probability of success π_i, respectively. Assume that equal sample sizes, n, are required.

Rule of Thumb

To compare two proportions, π_0 and π_1 use the formula

$$n = \frac{16\bar{\pi}(1-\bar{\pi})}{(\pi_0 - \pi_1)^2}\ , \tag{2.27}$$

where $\bar{\pi} = (\pi_0 + \pi_1)/2$ is used to calculate the average variance.

Illustration

For $\pi_0 = 0.3$ and $\pi_1 = 0.1$, $\bar{\pi} = 0.2$ so that the required sample size per group is $n = 64$.

Basis of the Rule

Use equation (2.3) with the variance estimated by $\bar{\pi}(1-\bar{\pi})$.

Discussion and Extensions

Some care should be taken with this approximation. It is reasonably good for values of n that come out between 10 and 100. For larger (or smaller) resulting sample sizes using this approximation, more exact formulae should be used. For more extreme values, use tables of exact values given by Haseman (1978) or use more exact formulae (see van Belle et al. 2003). Note that the tables by Haseman are for one-tailed tests of the hypotheses, thus they will tend to be smaller than sample sizes based on the two-tailed assumption in equation (2.3).

An upper limit on the required sample size is obtained using the maximum variance of 1/4 which occurs at $\pi_i = 1/2$. For these values $\sigma = 1/2$ and the sample size formula becomes

$$n = \frac{4}{(\pi_0 - \pi_1)^2}\ . \tag{2.28}$$

This formula produces a conservative estimate of the sample size. Using the specification in the illustration produces a sample size of $n = 4/(0.3 - 0.1)^2 = 100$–

considerably higher than the value of 64. This is due to larger value for the variance. This formula is going to work reasonably well when the proportions are centered around 0.5.

Why not use the variance stabilizing transformation for the binomial case? This has been done extensively. The variance stabilizing transformation for a binomial random variable, X, the number of successes in n Bernoulli trials with probability of success, π is,

$$Y = \sin^{-1}\sqrt{\frac{X}{n}}\ , \tag{2.29}$$

where the angle is measured in radians. The variance of $Y = 1/(4n)$. Using the square transformation in equation (2.3) gives

$$n = \frac{4}{\left(\sin^{-1}\sqrt{\pi_0} - \sin^{-1}\sqrt{\pi_1}\right)^2}\ . \tag{2.30}$$

For the example this produces $n = 60.1 = 61$; the value of n=64, using the more easily remembered equation (2.27) is compatible.

For proportions less than 0.05, $\sin^{-1}\sqrt{\pi} \approx \sqrt{\pi}$. This leads to the sample size formula,

$$n = \frac{4}{(\sqrt{\pi_0} - \sqrt{\pi_1})^2}\ , \tag{2.31}$$

which is linked to the Poisson formulation in equation (2.20).

Equation (2.1) assumes that the variances are equal or can be replaced by the average variance. Many reference books, for example Selvin (1996), Lachin (2000) and Fleiss et al. (2003) use equation (2.2) with the variances explicitly accounted for. The hypothesis testing situation is, $\mathrm{H}_0 : \pi_0 = \pi_1 = \pi$ and $\mathrm{H}_1 : \pi_0 \neq \pi_1$, say, $\pi_0 - \pi_1 = \delta$. This produces the fundamental equation

$$0 + z_{1-\alpha/2}\sqrt{\frac{2\pi(1-\pi)}{n}} = \delta - z_{1-\beta}\sqrt{\frac{\pi_0(1-\pi_0)}{n} + \frac{\pi_1(1-\pi_1)}{n}}\ . \tag{2.32}$$

Solving this equation for n produces

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\pi(1-\pi)} + z_{1-\beta}\sqrt{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)}\right)^2}{(\pi_0 - \pi_1)^2}\ . \tag{2.33}$$

Using this equation (2.33) for the data in the illustration with $\pi_0 = \pi_1 = 0.2$ under the null hypothesis, and $\pi_0 = 0.3, \pi_1 = 0.1$ under the alternative hypothesis produces a sample size of $n = 61.5 \sim 62$. Clearly, the approximation works very well.

As illustrated, equation (2.27) produces reasonable estimates of sample sizes for n in the range from 10 to 100. For smaller sampling situations exact tables should be used.

2.10 WHEN UNEQUAL SAMPLE SIZES MATTER; WHEN THEY DON'T

Introduction

In some cases it may be useful to have unequal sample sizes. For example, in epidemiological studies it may not be possible to get more cases, but more controls are available. Suppose n subjects are required per group, but only n_0 are available for one of the groups, assuming that $n_0 < n$. What is the number of subjects, kn_0, required in the second group in order to obtain the same precision as with n in each group?

Rule of Thumb

To get equal precision with a two-sample situation with n observations per sample given n_0 $(n_0 < n)$ in the first sample and kn_0 observations in the second sample, choose

$$k = \frac{n}{2n_0 - n} \ . \tag{2.34}$$

Illustration

Suppose that sample size calculations indicate that $n = 16$ cases and controls are needed in a case-control study. However, only 12 cases are available. How many controls will be needed to obtain the same precision? The answer is $k = 16/8 = 2$ so that 24 controls will be needed to obtain the same precision as with 16 cases and controls.

Basis of the Rule

For two independent samples of size n, the variance of the estimate of difference (assuming equal variances) is proportional to

$$\frac{1}{n} + \frac{1}{n} = \frac{2}{n} \ . \tag{2.35}$$

Given a sample size $n_0 < n$ available for the first sample and a sample size kn_0 for the second sample and then equating the variances for the two designs, produces

$$\frac{1}{n_0} + \frac{1}{kn_0} = \frac{2}{n} \ . \tag{2.36}$$

Solving for k produces the result.

Discussion and Extensions

The rule of thumb implies that there is a lower bound to the number of observations in the smaller sample size group. In the example the required precision, as measured

by the sum of reciprocals of the sample sizes is $1/8$. Assuming that $k = \infty$ requires 8 observations in the first group. This is the minimum number of observations. Another way of saying the result is that it is not possible to reduce the variance of the difference to less than $1/8$. This result is asymptotic. How quickly is the minimum value of the variance of the difference approached? This turns out the be a function of k only.

This approach can be generalized to situations where the variances are not equal. The derivations are simplest when one variance is fixed and the second variance is considered a multiple of the first variance (analogous to the sample size calculation).

Now consider two designs, one with n observations in each group and the other with n and kn observations in each group. The relative precision of these two designs is

$$\frac{SE_k}{SE_0} = \sqrt{\frac{1}{2}\left(1 + \frac{1}{k}\right)}, \tag{2.37}$$

where SE_k and SE_0 are the standard errors of the designs with kn and n subjects in the two groups, respectively. Using $k = 1$, results in the usual two-sample situation with equal sample size. If $k = \infty$, the relative precision is $\sqrt{0.5} = 0.71$. Hence, the best that can be done is to decrease the standard error of the difference by 29%. For $k = 4$ the value is already 0.79 so that from the point of view of precision there is no reason to go beyond four or five times more subjects in the second group than the first group. This will come close to the maximum possible precision in each group.

There is a converse to the above rule: minor deviations from equal sample sizes do not affect the precision materially. Returning to the illustration, suppose the sample size in one group is 17, and the other is 15 so that the total sampling effort is the same. In this case the precision is proportional to

$$\frac{1}{17} + \frac{1}{15} = 0.1255.$$

This compares with 0.125 under the equal sampling case. Thus the precision is virtually identical and some imbalance can be tolerated. Given that the total sampling effort remains fixed a surprisingly large imbalance can be tolerated. Specifically, if the samples are split into $0.8n$ and $1.2n$ the decrease in precision, as measured by the reciprocal of the sample sizes, in only 4%. So an imbalance of approximately 20% has a small effect on precision. A split of $0.5n, 1.5n$ results in a 33% reduction in precision. The results are even more comparable if the precision is expressed in terms of standard errors rather than variances.It must be stressed that this is true only if the total sampling effort remains the same. Also, if the variances are not equal the results will differ, but the principle remains valid. See the next rule for further discussion. Lachin (2000) gives a good general approach to sample size calculations with unequal numbers of observations in the samples.

The importance of the issue of unequal sample sizes must be considered from two points of view: when unequal sample sizes matter and when they don't. It matters when multiple samples can be obtained in one group. It does not matter under moderate degrees of imbalance.

2.11 DETERMINING SAMPLE SIZE WHEN THERE ARE DIFFERENT COSTS ASSOCIATED WITH THE TWO SAMPLES

Introduction

In some two-sample situations the cost per observation is not equal and the challenge then is to choose the sample sizes in such a way so as to minimize cost and maximize precision, or minimize the standard error of the difference (or, equivalently, minimize the variance of the difference). Suppose the cost per observation in the first sample is c_0 and in the second sample is c_1. How should the two sample sizes n_0 and n_1 be chosen?

Rule of Thumb

To minimize the total cost of a study, choose the ratio of the sample sizes according to

$$\frac{n_1}{n_0} = \sqrt{\frac{c_0}{c_1}} \ . \tag{2.38}$$

This is the square root rule: Pick sample sizes inversely proportional to square root of the cost of the observations. If costs are not too different, then equal sample sizes are suggested (because the square root of the ratio will be closer to 1).

Illustration

Suppose the cost per observation for the first sample is 160 and the cost per observation for the second sample is 40. Then the rule of thumb recommends taking twice as many observations in the second group as compared to the first. To calculate the specific sample sizes, suppose that on an equal sample basis 16 observations are needed. To get equal precision with n_0 and $2n_0$, use equation (2.36) with $k = 2$ to produce 12 and 24 observations, respectively. The total cost is then $12 \times 160 + 24 \times 40 = 2800$ compared with the equal sample size cost of $16 \times 160 + 16 \times 40 = 3200$ for a 10% saving in total cost. This is a modest saving in total cost for a substantial difference in cost per observation in the two samples. This suggests that costs are not going to play a major role in sample size determinations.

Basis of the Rule

The cost, C, of the experiment is

$$C = c_0 n_0 + c_1 n_1 \ , \tag{2.39}$$

where n_0 and n_1 are the number of observations in the two groups, respectively, and are to be chosen to minimize

$$\frac{1}{n_0} + \frac{1}{n_1} \ , \tag{2.40}$$

subject to the total cost being C. This is a linear programming problem with solutions:

$$n_0 = \frac{C}{c_0 + \sqrt{c_0 c_1}} \tag{2.41}$$

and

$$n_1 = \frac{C}{c_1 + \sqrt{c_0 c_1}} \ . \tag{2.42}$$

When ratios are taken, the result follows.

Discussion and Extensions

The argument is similar to that in connection with the unequal sample size rule of thumb. There are also two perspectives in terms of precision: when do costs matter, when do they not? The answer is in the same spirit as that associated with Rule 2.10. On the whole, costs are not that important. A little algebra shows that the total cost of a study under equal sample size, say C_{equal} is related to the total cost of a study using the square root rule, say C_{optimal} as follows,

$$\frac{C_{\text{equal}} - C_{\text{optimal}}}{C_{\text{equal}}} = \frac{1}{2} - \frac{\sqrt{c_0 c_1}}{c_0 + c_1} \ . \tag{2.43}$$

The following array displays the savings as a function of the differential costs per observation in the two samples. It is assumed that the costs are higher in the first sample.

$\frac{c_0}{c_1}$	1	2	5	10	15	20	100
$\frac{C_{\text{equal}} - C_{\text{optimal}}}{C_{\text{equal}}}$	0	3%	13%	21%	26%	29%	40%.

These results are sobering. The savings can never be greater than 50%. Even a five-fold difference in cost per observation in the two groups results in only a 13% reduction in the total cost. These results are valid for all sample sizes, that is, the percentage savings is a function of the different costs per observation, not the sample sizes. A similar conclusion is arrived at in Rule 6.5 on page (124).

This discussion assumes–unrealistically–that there are no overhead costs. If overhead costs are taken into account, the cost per observation will change and, ordinarily, reduce the impact of cost on precision.

The variance of an observation can be considered a cost; the larger the variance the more observations are needed. The discussions for this rule and the previous rule can be applied to differential variances.

Imbalance in sample sizes, costs, and variances can all be assessed by these rules. On the whole, minor imbalances have minimal effects on costs and precision. Ordinarily initial considerations in study design can ignore these aspects and focus on the key aspects of estimation and variability.

2.12 USE THE RULE OF THREES TO CALCULATE 95% UPPER BOUNDS WHEN THERE HAVE BEEN NO EVENTS

Introduction

The rule of threes can be used to address the following type of question: "I am told by my physician that I need a serious operation and have been informed that there has not been a fatal outcome in the 20 operations carried out by the physician. Does this information give me an estimate of the potential postoperative mortality?" The answer is "yes!"

Rule of Thumb

Given no observed events in n trials, a 95% upper bound on the rate of occurrence is

$$\frac{3}{n}\,. \tag{2.44}$$

Illustration

Given no observed events in 20 trials a 95% upper bound on the rate of occurrence is $3/20 = 0.15$. Hence, with no fatalities in 20 operations the rate could still be as high as 0.15 or 15%.

Basis of the Rule

Formally, assume Y is Poisson (θ) using n samples. The Poisson has the useful property that the sum of independent Poisson variables is also Poisson. Hence in this case, $Y_1 + Y_2 + ... + Y_n$ is Poisson $(n\theta)$ and the question of at least one Y_i not equal to zero is the probability that the sum, $\sum Y_i$, is greater than zero. Specify this probability to be, say, 0.95 so that

$$P(\sum Y_i = 0) = e^{-n\theta} = 0.05. \tag{2.45}$$

Taking logarithms, produces

$$n\theta = -\ln(0.05) = 2.996 \sim 3. \tag{2.46}$$

Solving for θ,

$$\theta = \frac{3}{n}\,. \tag{2.47}$$

This is one version of the "rule of threes."

Discussion and Extensions

The equation, $n\theta = 3$, was solved for θ. It could have solved it for n as well. To illustrate this approach, consider the following question: "The concentration of

Cryptosporidium in a water source is θ per liter. How many liters must I take to make sure that I have at least one organism?" The answer is, "Take $n = 3/\theta$ liters to be 95% certain that there is at least one organism in the sample."

Louis (1981) derived the rule of threes under the binomial model. He also points out that this value is the 95% Bayesian prediction interval using a uniform prior. For an interesting discussion see Hanley and Lippman-Hand (1983). For other applications van Belle et al. (2003).

The key to the use of this result is that the number of trials without observed adverse events is known. The results have wide applicability. A similar rule can be derived for situations where one or more events are observed, but it is not as interesting as this situation.

2.13 SAMPLE SIZE CALCULATIONS SHOULD BE BASED ON THE WAY THE DATA WILL BE ANALYZED

There are obviously many ways to do sample size calculations, although the simple ones discussed in this chapter predominate. As a first approximation, calculate sample sizes for pairwise comparisons. While there are formulae for more complicated situations, they require specific alternative hypotheses which may not be acceptable or meaningful to the researcher.

Rule of Thumb

Ordinarily the sample size calculation should be based on the statistics used in the analysis of the data.

Illustration

If a sample size calculation is based on the normal model–that is, based on continuous data–then the data should not be analyzed by dichotomizing the observations and doing a binomial test.

Basis of the Rule

One of the key ingredients in sample size calculations is power which is associated with a particular statistical procedure. Carrying out some other statistical procedure during the analysis may alter the anticipated power. In addition, the estimated treatment effect may no longer be meaningful in the scale of the new statistic.

Discussion and Extensions

This rule is often more honored in the breach. For example, sample size calculations may be based on the two sample *t*-test of two treatments in a four-treatment completely

randomized design, but an analysis of variance is actually carried out. This is probably the least offensive violation of the rule. The illustration above represents a more flagrant violation.

One situation where breaking the rule may make sense is where a more involved analysis will presumably increase the precision of the estimate and thus increase the power. This is a frequent strategy in grant applications where sample size calculations are required on short notice and the complexity of the data cannot be fully described at the time of application. For further discussion of this point see Friedman et al. (1998, page 99).

This chapter has shown the versatility of some simple formulae for calculating sample sizes. Equation (2.3) is the basis for many of the calculations. Equation (2.2) permits sample size calculations where the variances and the sample sizes are not equal. This approach needs only to be used when there are marked inequalities. For planning purposes it is best to start with the stronger assumption of equal sample sizes and equal variances.