Chapter 1
Calculating Sample Size in Anthropometry

Carine A. Bellera, Bethany J. Foster, and James A. Hanley

Abstract Sample size estimation is a fundamental step when designing clinical trials and epidemiological studies for which the primary objective is the estimation or the comparison of parameters. One may be interested in the prevalence of overweight children in a given population; however, the true prevalence will remain unknown and cannot be observed unless the whole population is studied. Statistical inference is the use of statistics and random sampling to make inferences concerning the true parameters of a population. By choosing a representative sample, inference based on the observed prevalence leads to an estimation of the true parameter. But how many subjects should be sampled to obtain an accurate estimate of the prevalence? Similarly, how many subjects should we sample to show that this parameter is different from some fixed value?

We first review basic statistical concepts including random variables, population and sample statistics, as well as probability distributions such as the binomial and normal distributions. Principles of point and interval estimation, as well as hypothesis testing, are presented. We consider several commonly used statistics: single proportions, differences between two proportions, single means, differences between two means, and reference limits. For each parameter, point estimators are presented as well as methods for constructing confidence intervals. We then review general methods for calculating sample sizes. We first consider precision-based estimation procedures, where the sample size is estimated as a function of the desired degree of precision. Next, although there is greater emphasis on precision-driven estimation procedures, we also briefly describe power-based estimation methods. This approach requires defining a priori the difference one wishes to detect, the desired significance level, and the desired power of the test. Sample size estimation procedures are presented for each parameter, and examples are systematically provided.

Abbreviations and Notations

N	Population size
n	Sample size
ME	Margin of error
ε	Precision

C.A. Bellera (✉)
Department of Clinical Epidemiology and Clinical Research, Institut Bergonié, Regional Comprehensive Cancer Center, 229 Cours de l'Argonne, 33076 Bordeaux, France
e-mail: bellera@bergonie.org

V.R. Preedy (ed.), *Handbook of Anthropometry: Physical Measures of Human Form in Health and Disease*, DOI 10.1007/978-1-4419-1788-1_1,

μ	Population mean
m	Sample mean
σ^2	Population variance
s^2	Sample variance
π	Population proportion
p	Sample proportion
H_0	Null hypothesis
H_A	Alternative hypothesis
α	Type I error rate
β	Type II error rate
z_p	$100p\%$ standard normal deviate
BMI	Body mass index
DBP	Diastolic blood pressure

1.1 Introduction

Sample size estimation is a fundamental step when designing clinical trials and epidemiological studies for which the primary objective is the estimation or the comparison of parameters. One may be interested in the prevalence of a given health condition, e.g. obesity, in a specific population; however, the true prevalence will always remain unknown and cannot be determined unless the whole population is observed. Statistical inference is the use of statistics and random sampling to make inferences concerning the true parameters of a population. By selecting a representative sample, inference based on the **observed** prevalence leads to an estimation of the **true** parameter. But how many subjects should be sampled to obtain an accurate estimate of the prevalence?

Sample size estimation can be either precision-based or power-based. In the first scenario, one is interested in estimating a parameter, such as a proportion, or a difference between two means, with a specific level of precision. On the other hand, one might only be interested in testing whether two parameters differ. The sample size will be estimated as a function of the size of the difference one wishes to detect as well as the degree of certainty one wishes to obtain.

To understand the process of sample size estimation, it is important to be familiar with basic statistical concepts. We first review statistical principles, as well as general concepts of statistical inference, including estimation and hypothesis testing. Methods for sample size estimation are presented for various parameters using precision-based and power-based approaches, although there is greater emphasis on precision-driven estimation procedures (Gardner and Altman 1986, 1988).

Most concepts presented in this chapter are available in introductory statistical textbooks (Altman et al. 2000; Armitage et al. 2002) and texts focusing on the methodology of clinical trials (Machin et al. 1997; Friedman et al. 1998; Sackett 2001; Piantadosi 2005). We refer the interested readers to these works.

1.2 Basic Statistical Concepts

1.2.1 Random Variable

A random variable assigns a value to each subject of a population, such as weight, hair colour, etc. By random, it is implied that the true value of the variable cannot be known until it is observed. A variable

(for simplicity, we will often discard the term *random* throughout the rest of this chapter) is either **quantitative** or **qualitative**.

A quantitative variable is one that can be measured and can take a range of values, for example, waist circumference, size, age, or the number of children in a household. Quantitative variables include **discrete** and **continuous** variables. A discrete variable is one that can take only a limited range of values, or similarly, the possible values are distinct and separated, such as the number of children in a household. On the other hand, a continuous variable can take an infinite range of values, or similarly, can assume a continuous uninterrupted range of values, such as height or age.

A qualitative or categorical variable is one that cannot be numerically measured, such as the presence or absence of a disease, gender, or the colour of hair. A **dichotomous** or binary variable is one that can take one of two values, such as the presence or absence of a trait or state, or whether or not one is overweight.

1.2.2 Population Versus Sample Statistics

Suppose we are interested in describing the size of 10-year old girls attending English schools. Height in this population can be summarized by various quantities, such as the mean, the median or the variance. These quantities are called **population statistics** and are usually represented with Greek letters. Unless all 10-year old girls attending English schools are measured, the true value of population statistics, such as the mean height in our example, cannot be observed and is **unknown**. It is however possible to **estimate** the true value with some degree of certainty. This involves randomly sampling from the whole population of interest. Based on a random sample of 10-year old girls attending English schools, one observes the distribution of heights in this sample and calculates the observed mean. Quantities derived from an **observed sample** are called **sample statistics**, and are usually denoted using Roman letters.

Two random samples of equal size will usually not yield the same value of the sample statistic. The possible differences between the estimates from all possible samples (conceptual), or between each possible estimate and the true value are referred to as **sampling variation**. As a result, it is not possible to conclude that the observed sample mean corresponds to the true population mean. By using appropriate statistical methods, sample statistics can be used to make inferences about population statistics. In the next section, we present commonly used statistics.

1.2.3 Summarizing Data

1.2.3.1 Categorical Variables

Summarizing categorical variables involves counting the number of observations for each category of the variable. These counts are usually referred to as frequencies. The proportion of such counts among the total can also be represented.

1.2.3.2 Quantitative Variables

Continuous variables can be summarized using measures of **location** and **dispersion**. Measures of location, such as the mean or the median, represent the central tendency of distributions. Dispersion measures, such as the variance, represent the repartition of a variable around the central tendency.

Given a population of size N and a variable X with observed values $x_1,\ldots,x_N$, the population mean is given by: $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$. If only a random sample of size n is available, the sample mean m is calculated similarly and given by $m = \sum_{i=1}^{n} \frac{x_i}{n}$.

If observations are ordered in increasing order, the median is the middle observation of the sample. If the number of observations in a sample is odd, the median is the value of the $\frac{1}{2}(n+1)^{\text{th}}$ observation of the ordered sample, while it is the mean of the values of the two middle observations if the number of observations is even.

The most common measure of dispersion is the variance, or its square root, the standard deviation. Given a population of size N, the variance σ^2 is given by $\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{N}$, where μ is the population mean. The population standard deviation is $\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$. If a sample of size n is available, the sample variance s^2 is provided by $s^2 = \frac{\sum_i (x_i - m)^2}{n-1}$, where m is the sample mean. The denominator is slightly different from that of a population variance. This correction ensures that the parameter s^2 is an unbiased estimator of the population variance σ^2. Similarly, the sample standard deviation, usually denoted as SD, is calculated as $\text{SD} = \sqrt{\frac{\sum_i (x_i - m)^2}{n-1}}$.

Other measures of dispersion include the range, the inter-quartile range and reference limits. The range corresponds to the difference between the maximum and the minimum values. When in increasing order, the first or lower quartile corresponds to the value below which 25% of the data fall. The third or upper quartile corresponds to the value below which 75% of the data fall. There are several methods to compute quartiles: one involves calculating the first and third quartiles as the rank of the $\frac{1}{4}(n+1)$ th and $\frac{3}{4}(n+1)$ th observed values. Other methods are available, but will usually lead to relatively close results (Armitage et al. 2002).

The inter-quartile range is the difference between the upper and lower quartiles. Note that the second or middle quartile corresponds to the value below which 50% of the data fall, and as such, is equivalent to the median.

More generally, the $100p\%$ reference limit, where $0 < p < 1$, is the value below which $100p\%$ of the values fall. For example, the median is equivalent to the 50% reference limit. Reference limits are also called reference values, percentiles, or quantiles.

Example: A random sample of ten 20-year-old women leads to the following observed weights (in kg): 50, 55, 60, 61, 45, 52, 62, 54, 48, 53. The variable of interest X is the weight, which is a continuous variable. We first reorder this random sequence of $n = 10$ observations: 45, 48, 50, 52, 53, 54, 55, 60, 61, 62. Based on previous formulae, we have the following results:

- The sample mean is calculated as $m = \sum_{i=1}^{10} \frac{x_i}{10} = \frac{45 + \cdots + 62}{10} = 54$ kg
- The sample median is calculated as $\text{med} = \frac{53 + 54}{2} = 53.5$ kg

Table 1.1 Sample probability distribution for the categorical variable X defined as the number of copies of allele A

x	0	1	2
$P(X = x)$	3/10	5/10	2/10

- The sample variance is calculated as $s^2 = \dfrac{\sum_i (x_i - 54)^2}{10-1} = \dfrac{(45-54)^2 + \cdots + (62-54)^2}{9} = 32\ \text{kg}^2$
- The range is given by 62–45 = 17 kg

1.2.4 Probability Distributions

A distribution is defined as the set of frequencies of the values or categories of a measurement made on a group of persons. The distribution tells us either how many or what proportion of the group was found to have each value (or each range of values) out of all of the possible values that the quantitative measure can have (Last 2001).

Consider the variable X corresponding to the number of copies of a certain allele A. The variable X can take three distinct values: 0, 1, 2. Ten subjects are randomly selected from the general population and the following series of outcomes is observed: 0, 0, 2, 1, 1, 2, 1, 0, 1, 1. This series is called a random series, since previous values cannot be used to predict future observation. One can then calculate the proportion of subjects carrying two copies of the allele. This observed proportion equals 2/10 in our example. If subjects are randomly sampled indefinitely, this proportion will tend towards a limiting value, called the **probability** of carrying two alleles, denoted by $P(X = 2) = \pi$. The sample probability distribution of the variable X is represented in Table 1.1.

We have the following result: $\sum P(X = x) = 1$. This leads us to a fundamental property of probability distributions: for a given variable, the sum of the probabilities of each possible event equals 1.

1.2.4.1 Bernoulli and Binomial Distributions

Consider a dichotomous random variable X whose values can be either one state or the other, or the absence or presence of a specific trait. The true (unknown) proportion of the population that is in the index category of the state or trait (the probability that a randomly selected individual would be in this category ($X = 1$)) is denoted as π and the probability of being in the other (reference) category ($X = 0$) is thus $1 - \pi$. The parameter π defines the probability distribution of X, and is called the **Bernoulli distribution** (after J Bernoulli, a Swiss mathematician). The variable X is said to follow a Bernoulli distribution with parameter π. If, as here, the variable X is coded as 0 (reference category) or 1 (index category), its mean and variance are given respectively by $\text{mean}(X) = \pi$ and $\text{var}(X) = \pi(1-\pi)$.

Define the random variable Y as the observed number of subjects with a particular state of interest out of n randomly selected subjects. If the probability that this state is present is π, the variable Y, that is, the sum of Bernoulli random variables with parameter π, is said to follow a **binomial distribution** with mean and variance given respectively by $\text{mean}(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$. The probability that this state is present for k out of n subjects is given by: $P(Y = k) = C_n^k \pi^k (1-\pi)^{n-k} = \dfrac{n!}{(n-k)!k!}\pi^k(1-\pi)^{n-k}$.

Table 1.2 Probability distribution for the categorical variable *Y* defined as the number of subjects with state *S* out of 3 randomly selected subjects

y	0	1	2	3
$P(Y = y)$	1/8	3/8	3/8	1/8

When n is large, the distribution of Y will converge to a normal distribution with mean $n\pi$ and variance $n\pi(1-\pi)$.

Example: The presence of a health state S is assessed in $n = 3$ randomly selected subjects. The variable Y corresponds to the number of subjects with state S out of the $n = 3$ subjects, and can thus have 4 possible values: 0, 1, 2, or 3. Assume that the state S is present in $\pi = 50\%$ of the population. The probability of observing $k = 2$ subjects with state S out of $n = 3$ randomly selected subjects is given by: $P(Y=2) = C_3^2\pi^2(1-\pi)^{3-2} = \frac{3!}{(3-2)!2!}(0.50)^2(0.50)^{3-2} = \frac{3}{8}$.

The complete probability distribution of Y can be computed and is represented in Table 1.2.

1.2.4.2 The Normal Distribution

Consider a continuous random variable with values that vary between $-\infty$ and $+\infty$. The most important continuous probability distribution is the **Gaussian distribution** (after Karl Gauss, a German mathematician), also called the **normal distribution**, which is indexed by two parameters: the mean and the variance. If X follows a normal distribution with mean μ and variance σ^2, we use the following notation: $X \sim N(\mu,\sigma^2)$.The curve representing the normal distribution is symmetric, so that the mean and the median fall at the centre of the symmetry. It is described by the function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. The probability that a random variable falls between two values, A and B, is represented graphically by the area under the curve of the probability distribution and between the two vertical lines with coordinates $x = A$ and $x = B$, as illustrated in Fig. 1.1.

Numerically, this probability is obtained by computing the integral $P(A < x < B) = \int_A^B f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$. Calculation of this integral is not feasible using simple calculation tools; however, the properties of the normal distribution simplify the probability estimation in some cases. For example, the probability that any random normal variable $X \sim N(\mu,\sigma^2)$ falls within one standard deviation of the mean is known to be about 68%, that is $P(\mu - \sigma < x < \mu + \sigma) = 68\%$, as presented in Fig. 1.2.

Similarly, about 95% of the distribution falls within two standard deviations of the mean ($P(\mu - 2\sigma < x < \mu + 2\sigma) = 95\%$), and 99.7% within three standard deviations ($P(\mu - 3\sigma < x < \mu + 3\sigma) = 99.7\%$). For other values of the normal distribution, one has to either calculate the integral $P(A < x < B) = \int_A^B f(x)dx$ (by hand or using a mathematical or statistical software) or rely on statistical tables for which these probabilities are tabulated for values of μ and σ. There is however an infinite number of values for these parameters, and it is therefore not possible to have tables for all of them. Interestingly, every normal distribution with parameters μ and σ can be expressed in terms of a normal distribution with mean 0 and variance 1, called the **standard normal distribution**. Indeed, if $X \sim N(\mu,\sigma^2)$, then one can define the variable Z such that $Z = \frac{X-\mu}{\sigma}$. It can be shown that $Z \sim N(0,1)$.

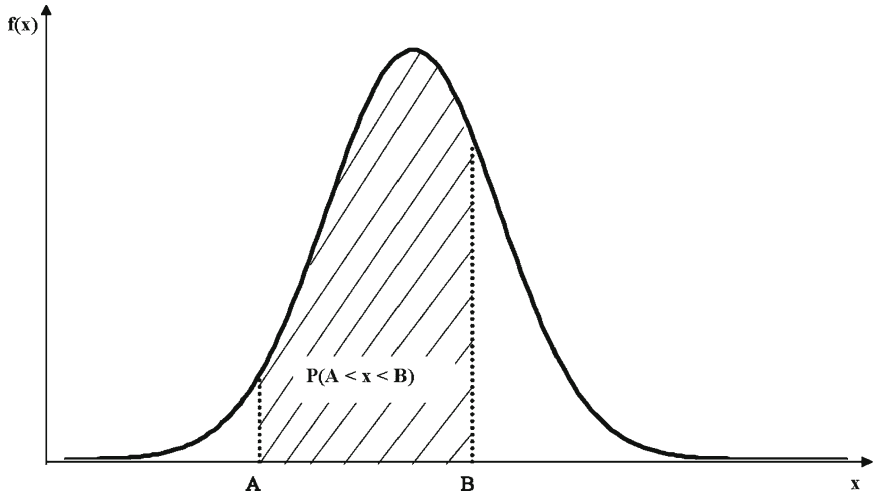

Fig. 1.1 Normal probability distribution function for a variable *X*. The *shaded area* represents the probability that the variable *X* falls between *A* and *B*. Numerically, this probability is obtained by computing the integral $P(A < x < B) = \int_A^B f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

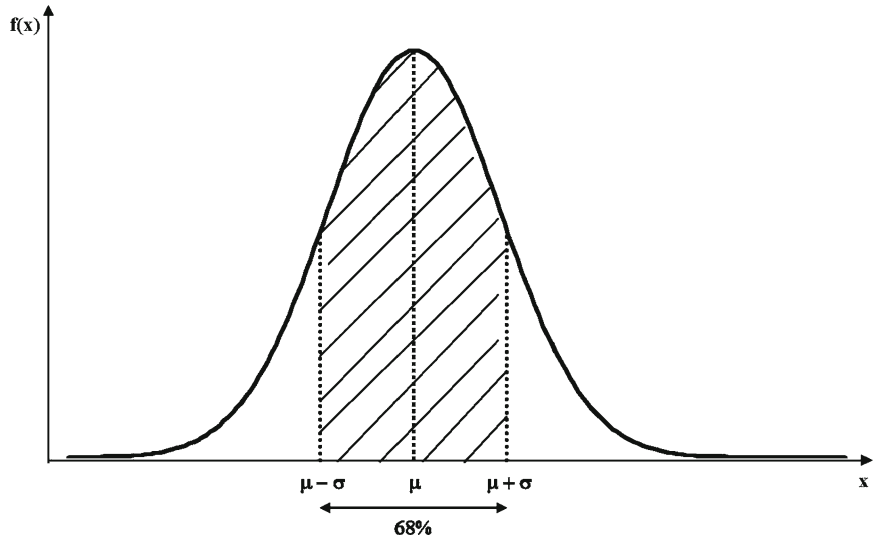

Fig. 1.2 Probability distribution function for a normal random variable *X* with mean μ and variance σ^2. The *shaded area* represents the probability that the normal variable *X* falls within the interval from the mean minus one standard deviation to the mean plus one standard deviation, that is $P(\mu - \sigma < x < \mu + \sigma)$, which is about 68%

Thus, computing the probability that a variable $X \sim N(\mu,\sigma^2)$ belongs to the interval [A;B], is equivalent to computing the probability that a variable $Z \sim N(0,1)$ belongs to the interval $\left[\frac{A-\mu}{\sigma}; \frac{B-\mu}{\sigma}\right]$. Tables of the standard normal distribution are available in most statistical textbooks. Important tables are associated with the standard normal distribution, including the table $P(z)$, which provides for each

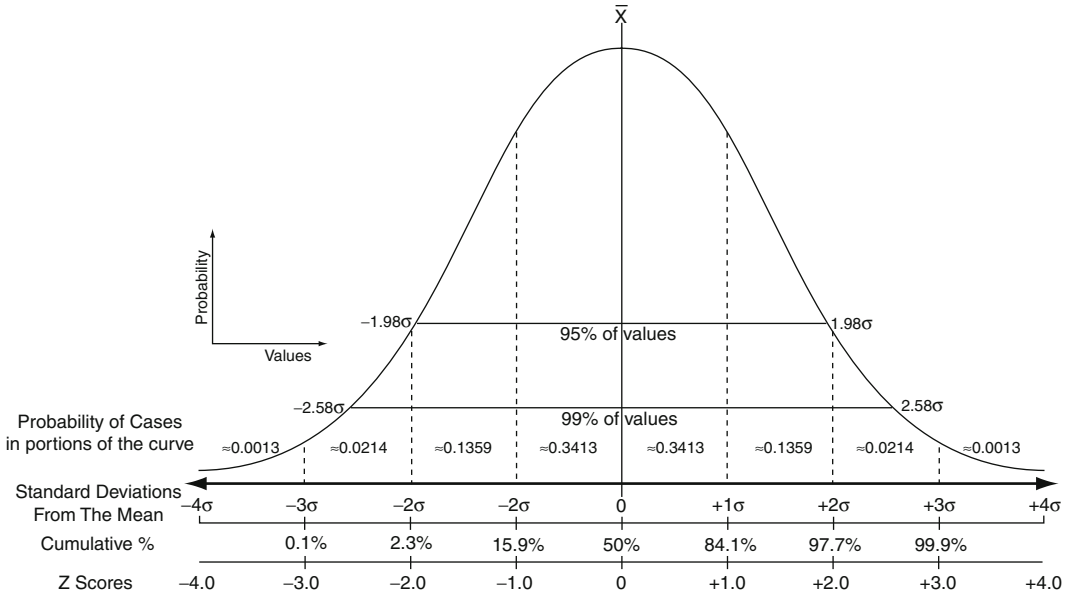

Fig. 1.3 The normal distribution. This plot illustrates standard results of the normal distribution including standard deviations from the means and the *Z*-scores

value of z, the probability that Z falls outside the interval $[-z; z]$: $P(z) = P(Z < -z \text{ or } Z > z) = 1 - P(-z < Z < z)$. Some standard normal values are commonly used: $P(1.64) = 0.10$ and $P(1.96) = 0.05$, that is, $P(-1.96 < Z < 1.96) = 95\%$. Thus, if a standard normal variable is randomly selected, there is a 95% chance that its value will fall within the interval [–1.96; 1.96]. Equivalently, the probability that a standard normal random variable $Z \sim N(0,1)$ falls outside the interval [–1.96; 1.96] is 5%. Given the symmetry of the normal distribution, $P(Z > 1.96) = P(Z < -1.96) = 2.5\%$. These standard results are illustrated in Fig. 1.3. The *Z*-values are usually referred to as the *Z*-scores and are commonly used in anthropometry. They are discussed in greater detail in a subsequent chapter.

The normal distribution is commonly used in anthropometry, in particular to construct reference ranges or intervals. This allows one to detect measurements which are extreme and possibly abnormal. A typical example is the construction of growth curves (WHO Child growth standards 2006). In practice, however, data can be skewed (i.e. not symmetric) and thus observations do not follow a normal distribution (Elveback et al. 1970). In such cases, a transformation of the observations, such as logarithmic, can remove or at least reduce the skewness of the data (Harris and Boyd 1995; Wright and Royston 1999).

1.3 Principles of Statistical Estimation

There are two estimation procedures: **point estimation** and **interval estimation**. Point estimation provides a value that we hope to be as close as possible to the true unknown parameter value. Interval estimation provides an interval that has a fixed a priori probability of containing the true parameter value.

1.3.1 Point Estimation

1.3.1.1 Point Estimator and Point Estimate

Point estimation is the process of assigning a value to a population parameter based on the observation of a sample drawn from this population. The resulting numerical value is called the **point estimate**; the mathematical formula/function used to obtain this value is called the **point estimator**. While the point estimator is identical whatever the sample, the point estimate varies across samples.

1.3.1.2 Point Estimation for a Proportion

The variable of interest is binary, such as the presence or absence of a specific health condition. We are interested in π, the true prevalence of this condition. Assume we randomly draw a sample of n subjects and denote the observed number of subjects with the condition of interest by k. The point estimator of π is given by $p = \frac{k}{n}$, while point estimates correspond to the numerical values $p_1, p_2, \ldots,$ obtained from distinct samples of n observations.

1.3.1.3 Point Estimation for a Mean and a Variance

X is a continuous variable with population mean and variance denoted respectively by μ and σ^2. If n subjects are randomly selected with subjects i=1 to n having respectively observed values x_1 to x_n, the mean μ, the variance σ^2 and the standard deviation SD are estimated respectively by $m = \sum_{i=1}^{n} \frac{x_i}{n}$, $s^2 = \sum_{i=1}^{n} \frac{(x_i - m)^2}{n-1}$ and $\mathrm{SD} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - m)^2}{n-1}}$.

1.3.1.4 Point Estimation for a Reference Limit

Let $X_1, X_2, \ldots, X_n$ be the measurements for a random sample of n individuals. The $100p\%$ reference limit, where $0 < p < 1$, is the value below which $100p\%$ of the values fall. Reference limits may be estimated using nonparametric or parametric (i.e. distribution-based) approaches (Wright and Royston 1999).

The simplest approach is to find the empirical reference limit, based on the order statistics. The k^{th} order statistic, denoted as $X_{(k)}$, of a statistical sample is equal to its k$^{\text{th}}$-smallest value. Thus, given our initial sample $X_1, X_2, \ldots, X_n$, the order statistics are $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, and represent the observations with first, second, …, n^{th} smallest value, respectively. An empirical estimation of the $100p\%$ reference limit is given by the value which has rank $[p(n+1)]$, where [.] denotes the nearest integer. For example, for a sample of size $n = 199$ and $p = 0.025$, $[p(n+1)]=5$, and thus the value of the fifth order statistic, $X_{(5)}$, provides a point estimate of the 2.5% reference limit.

Reference limits can also be estimated based on parametric methods. Assume that the observations follow a normal distribution with mean μ and variance σ^2. For a random sample of size n, the sample mean and sample standard deviation are given by m and SD, respectively. The $100p\%$ reference limit is then estimated as $m + z_p \mathrm{SD}$, where z_p is the $100p\%$ standard normal deviate.

1.3.2 Interval Estimation

1.3.2.1 Principles

A single-value point estimate, without any indication of its variability, is of limited value. Because of sample variation, the value of the point estimate will vary across samples, and it will not provide any information regarding precision. One could provide an estimation of the variance of the point estimator. However, it is more common to provide a range of possible values. A **confidence interval** has a specified probability of containing the parameter value (Armitage et al. 2002). By definition, this probability, called the coverage probability, is $(1-\alpha)$, where $0<\alpha<1$. If one selects a sample of size n at random, using say a set of random numbers, then the sample that one gets to observe is just one sample from among a large number of possible samples one might have observed, had the play of chance been otherwise. A different set of n random numbers would lead to a different possible sample of the same size n and a different point estimate of the parameter of interest, along with its confidence interval. Of the many possible interval estimates, some $100(1-\alpha)\%$ of these intervals will capture the true value. This implies that we cannot be sure that the $100(1-\alpha)\%$ confidence interval calculated from the actual sample will capture the true value of the parameter of interest. The computed interval might be one of the possible intervals (with proportion $100\alpha\%$) that do not contain the true value. The most commonly used coverage probability is 0.95, that is $\alpha = 0.05 = 5\%$, in which case the interval is called a 95% confidence interval.

1.3.2.2 Interval Estimation for a Proportion

In this situation, the variable of interest is binary, such as the presence or absence of a specific health condition, and we are interested in π, the prevalence rate of this condition. Given a random sample of n subjects, a point estimator for the proportion π is given by $p = \frac{k}{n}$, where k is the observed number of subjects with the condition of interest in the sample. When n is large, the sampling distribution of p is approximately normal, with mean π and variance $\frac{\pi(1-\pi)}{n}$. Thus, a $100(1-\alpha)\%$ confidence interval for π is given by $\left[p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right]$. In smaller samples ($np < 5$ or $n(1-p)<5$), exact methods based on the binomial distribution should be applied rather than a normal approximation to it (Machin et al. 1997).

Example: Assume one is interested in estimating π, the proportion of overweight children. In a selected random sample of $n = 100$ children, the observed proportion of overweight children is $p = 40\%$.A 95% confidence interval $\left(z_{1-\alpha/2} = 1.96\right)$ for the true proportion π of overweight children is given by $\left[0.40 \pm 1.96\sqrt{\frac{0.40(0.60)}{100}} \right]$, that is [30%; 50%].

1.3.2.3 Interval Estimation for the Difference Between Two Proportions

The variables of interest are binary, and we are interested in π_1 and π_2, the prevalence rates of a specific condition in two independent samples of size n_1 and n_2. Let k_1 and k_2 denote the number of

observed subjects with the condition of interest in these two samples. A point estimator for the difference $\pi_1 - \pi_2$ is given by $p_1 - p_2 = \frac{k_1}{n_1} - \frac{k_2}{n_2}$, where p_1 and p_2 are the two sample proportions. When n_1 and n_2 are large, the sampling distribution of $p_1 - p_2$ is approximately normal with mean $\pi_1 - \pi_2$ and variance $\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$. Thus, a 100(1 − α)% confidence interval for the true difference in proportions $(\pi_1 - \pi_2)$ is given by $\left[(p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right]$. In case of smaller samples ($n_1 p_1 < 5$ or $n_1(1 - p_1) < 5$ or $n_2 p_2 < 5$ or $n_2(1 - p_2) < 5$), exact methods based on the binomial distribution should be applied rather than a normal approximation to it (Machin et al. 1997).

Example: Assume a randomized trial is conducted to compare two physical activity interventions for reducing waist circumference. A total of 110 and 100 subjects are assigned to receive intervention A or B, respectively. The observed success rate is 40% in group A and 20% in group B. A 95% confidence interval for the true difference in success rates is given by

$$\left[(0.40 - 0.20) \pm 1.96\sqrt{\frac{0.40 \times 0.60}{110} + \frac{0.20 \times 0.80}{100}}\right],$$ that is, [8%; 32%].

1.3.2.4 Interval Estimation for a Mean

The variable of interest is continuous with true mean μ and true variance σ^2. We are interested in estimating the true population mean μ. A total of n subjects have been randomly selected from this population of interest. When n is large ($n > 30$), the distribution of the sample mean m is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$. A 100(1–α) confidence interval for μ is thus given by $\left[m \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right]$. One can use s^2 as an estimate of the population variance, unless the variance σ^2 is known. If the sample size is small, and if the variable of interest is known to be normal, then a 100(1–α) confidence interval for μ is given by $\left[m \pm t_{n-1,\alpha}\sqrt{\frac{s^2}{n}}\right]$, where $t_{n-1,\alpha}$ corresponds to the α% tabulated point of the $t_{n-1,\alpha}$ distribution.

Example: One is interested in estimating the average weight of 15-year-old girls. After selecting a random sample of 100 girls, the observed average weight is 55 kg, and the standard deviation is 15 kg. A 95% confidence interval for the mean weight of 15-year-old girls is thus given by: $\left[55 \pm 1.96\sqrt{\frac{15^2}{100}}\right]$, that is [52; 58].

1.3.2.5 Interval Estimation for the Difference Between Two Means

Given two independent samples of size n_1 and n_2 with respective means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, a point estimator for the difference $\mu_1 - \mu_2$ is given $m_1 - m_2$, where m_1 and m_2 correspond

to the two sample means. When n_1 and n_2 are large, the distribution of m_1–m_2 is approximately normal with mean μ_1–μ_2 and variance $\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$. A 100(1–$\alpha$)% confidence interval for the true difference in means is then given by $\left[(m_1-m_2)\pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$. Unless the variances σ_1^2 and σ_2^2 are known, one can use the sample variances s_1^2 and s_2^2 as their respective estimates: $\left[(m_1-m_2)\pm z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$.

Example: One wishes to compare the diastolic blood pressure (DBP) between two populations of subjects A and B. In a sample of 50 randomly selected subjects from population A, the observed DBP mean is 110 mmHg, while the mean DBP in 60 patients randomly selected from population B is about 100 mmHg. The observed standard deviations are, respectively, 10 mmHg and 8 mmHg. A 95% confidence interval for the true difference in DBP means is thus $\left[(110-100)\pm 1.96\times\sqrt{\frac{10^2}{50}+\frac{8^2}{60}}\right]$, that is, [6.6; 13.4].

1.3.2.6 Interval Estimation for a Reference Limit

Using a nonparametric approach, a confidence interval for a reference limit can be expressed in terms of order statistics (Harris and Boyd 1995). An approximate 100(1–α)% confidence interval for the 100p% reference limit is given by the order statistics $x_{(r)}$ and $x_{(s)}$, where r is the largest integer less than or equal to $np+\frac{1}{2}-z_{\frac{\alpha}{2}}\sqrt{np(1-p)}$, and s is the smallest integer greater than or equal to $np+\frac{1}{2}+z_{\frac{\alpha}{2}}\sqrt{np(1-p)}$.

Example: In their example, Harris and Boyd are interested in the 90% confidence interval $\left(z_{1-\alpha/2}=1.645\right)$ for the 2.5% (p=0.025) reference limit based on a random sample of size n=240 (Harris and Boyd 1995). The lower bound of this interval corresponds to the value of the r^{th} order statistic, where r is the largest integer less than or equal to $np+\frac{1}{2}-z_{\frac{\alpha}{2}}\sqrt{np(1-p)}=240\times 0.025+\frac{1}{2}-1.645\sqrt{100\times 0.025\times 0.975}$, that is, 2. Similarly, the upper bound corresponds to the value of the s^{th} order statistic, where s is the smallest integer greater than or equal to $np+\frac{1}{2}+z_{\frac{\alpha}{2}}\sqrt{np(1-p)}$, that is, 11. Referring to the ordered observations, bounds of the 90% confidence interval for the 2.5% reference limit are thus provided by the values of the second and eleventh order statistics.

Confidence intervals can also be built using parametric methods. Assume that the observations follow a normal distribution with mean μ and variance σ^2. For a random sample of size n, the sample mean and sample standard deviation are given by m and SD, respectively. The parametric estimator of the $100p\%$ reference limit is $m+z_p\text{SD}$, where z_p is the $100p\%$ standard normal deviate. The variance of this estimator is given by $\frac{\sigma^2}{n}\left(1+\frac{z_p^2}{2}\right)$. A $100(1-\alpha)\%$ confidence interval for the $100p\%$ reference limit is thus $\left[m+z_p\text{SD}\pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\text{SD}^2}{n}\left(1+\frac{z_p^2}{2}\right)}\right]$. See Harris and Boyd for worked examples (Harris and Boyd 1995).

1.4 Precision-Based Sample Size Estimation

Sample size can be computed using either a precision-based or power-based approach. In anthropometry, however, there is greater emphasis on precision-driven estimation procedures (Gardner and Altman 1986, 1988).

For any given parameter, the length of the confidence interval is a function of the sample size. More specifically, the larger the sample, the narrower the confidence interval. Thus, the main purpose of confidence intervals is to indicate the (im)precision of the sample study estimates as population values (Gardner and Altman 1988). Conversely, one can fix the desired length of the interval and estimate the number of subjects needed accordingly. Formulae are available to estimate the sample size as a function of the precision, which can be expressed in two ways.

One can express the length of the interval in absolute terms based on the **absolute margin of error**, ME, which represents half the width of the confidence interval, and is the quantity often quoted as the "plus or minus" in lay reports of surveys. In such case, the confidence interval is expressed as "estimate ± ME". It is also possible to express the length of the interval in **relative** terms based on the **precision** usually denoted by ε. In such cases, the confidence interval is expressed as "estimate $\pm\varepsilon\times$ estimate". There is a direct relationship between the margin of error and the precision since ME $=\varepsilon\times$ estimate; sample size can thus be estimated as a function of either parameter.

Note that the term *error* is often a source of confusion when dealing with sample size and power calculations. Therefore, it is always very important to provide a precise definition of this term, that is, to clarify whether we are referring to an absolute or relative error.

Throughout this section, we consider that observations from the same sample are independent. In the case of two-sample problems, we consider that the two samples are independent.

1.4.1 Dichotomous Variables

1.4.1.1 Sample Size for Estimating a Single Proportion with a Given Precision

A $100(1-\alpha)\%$ confidence interval for the true proportion π given a sample size n and an anticipated value p is given by: $\left[p\pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$. The width of the confidence interval is thus given by

$2 \times z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$ and depends on the number of subjects in the sample. Conversely, the number of subjects needed to estimate the proportion will depend on the degree of precision (indicated by the maximum width of the confidence interval) one is willing to accept. The margin of error, ME, is half the width of the confidence interval. If one wants the absolute margin of error to be at most ME, that is, one wants $z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq \text{ME}$, then the number of subjects should be at least $n = \frac{z^2_{1-\frac{\alpha}{2}} p(1-p)}{\text{ME}^2}$.

This problem can also be expressed in terms of precision (or relative error). The confidence interval $\left[p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right]$ can be rewritten as $[p \pm \varepsilon p]$, implying that the estimation of π is provided to within $100\varepsilon\%$ of its anticipated value, where ε is the precision of the estimation. Since ME = $p\varepsilon$,the required sample size can thus be expressed as $n = \frac{z^2_{1-\frac{\alpha}{2}} (1-p)}{p\varepsilon^2}$. Note that the sample size is maximized for $p = 0.50$.

If either np or $n(1-p)$ are small (i.e. below 5), exact approximations based on the binomial distribution should be applied rather than a normal approximation to it (Machin et al. 1997).

Example: One is interested in estimating the prevalence of overweight children with an absolute margin of error smaller than 0.05 (ME = 0.05). If the prevalence, π, is expected to be around 50% ($p = 0.50$), and interest is in estimating a 95% confidence interval $\left(z_{1-\frac{\alpha}{2}} = 1.96 \right)$, the minimum number of subjects needed should be $\frac{1.96^2 0.5(1-0.5)}{0.05^2}$, that is, at least 384 subjects.

1.4.1.2 Sample Size for Estimating the Difference Between Two Proportions with a Given Precision

Given two proportions π_1 and π_2 for two independent samples of size n_1 and n_2, a $100(1-\alpha)\%$ confidence interval for the true difference in proportions $(\pi_1 - \pi_2)$ is estimated by $\left[(p_1 - p_2) \pm z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right]$, where p_1 and p_2 are the observed sample proportions.

The width of the confidence interval is given by $2 \times z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Assuming samples are of equal size $n = n_1 = n_2$ and an absolute margin of error ME, the number of subjects in each sample should be at least $n = \frac{z^2_{1-\alpha/2} \left(p_1(1-p_1) + p_2(1-p_2) \right)}{\text{ME}^2}$. In terms of precision ε or relative error, where ME= $\varepsilon(p_1 - p_2)$, the required sample size is expressed as $n = \frac{z^2_{1-\alpha/2} \left(p_1(1-p_1) + p_2(1-p_2) \right)}{\varepsilon^2 (p_1 - p_2)^2}$. Similar formulae have been derived for the case of unequal sample sizes (Machin et al. 1997).

***Example*:** A study is set up to compare a new physical activity intervention to a standard one for overweight subjects. The aim is to reduce the body mass index (BMI) down to 30 cm/kg^2 or lower. The anticipated success rates are expected to be approximately $p_1 = 10\%$ and $p_1 = 25\%$ for the new and standard interventions respectively. The investigator would like to recruit two groups of patients with equal sample sizes $n = n_1 = n_2$, and provide a 95% confidence interval for difference in success rates with an absolute margin of error ME = 10%. The minimum number of subjects per group should be at least $n = \frac{1.96^2\left(0.10\times 0.90 + 0.25\times 0.75\right)}{0.10^2}$, that is, 107 subjects per group.

1.4.2 Continuous Variables

1.4.2.1 Sample Size for Estimating a Single Mean with a Given Precision

A 100(1–α)% confidence interval for a mean μ given anticipated mean m and assuming a large sample is given by $\left[m \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$. The width of the confidence interval is $2\times z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. If we want the absolute margin of error to be at most ME, then the number of subjects has to be greater than $n = \frac{z^2_{1-\frac{\alpha}{2}}\sigma^2}{\mathrm{ME}^2}$. If one wishes to express the sample size in terms of precision ε, where ME = εm, the required sample size is expressed as $n = \frac{z^2_{1-\frac{\alpha}{2}}\sigma^2}{\varepsilon^2 m^2}$. Unless the variance σ^2 is known, one can use a literature-based or experience-based s^2 as an estimate of the population variance.

Example: An investigator is interested in estimating diastolic blood pressure (DBP) in a specific population. The mean DBP and standard deviation are anticipated to be about 105 mmHg ($m = 105$) and 20 mmHg (SD = 20). If the desired relative precision is 5% ($\varepsilon = 5\%$), the required sample size for estimating a 95% confidence interval for the mean DBP should be at least $n = \frac{1.96^2 20^2}{0.05^2 105^2}$, that is 56 subjects.

1.4.2.2 Sample Size for Estimating the Difference Between Two Means with a Given Precision

Assuming independent samples of size n_1 and n_2, with respective sample means m_1 and m_2 and common variance σ^2, a 100(1–α)% confidence interval for the true difference in means is given by $\left[(m_1 - m_2) \pm z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$. The width of the confidence interval is thus given by $2\times z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$. Assuming samples of equal size $n = n_1 = n_2$ and a margin of error ME, the number of subjects in each sample should be at least $n = \frac{2\, z^2_{1-\alpha/2}\sigma^2}{\mathrm{ME}^2}$. If one wishes to express the

sample size in terms of precision ε, where ME = $\varepsilon(m_1 - m_2)$, the required sample size is expressed as $n = \frac{2\, z^2_{1-\alpha/2}\, \sigma^2}{\varepsilon^2 (m_1 - m_2)^2}$. Unless the variance σ^2 is known, one can use a literature-based or experience-based s^2 as an estimate of the population variance. Similar formulae have been derived for the case of unequal sample sizes (Machin et al. 1997).

Example: An investigator is interested in evaluating two treatments (A and B) aimed at decreasing cholesterol level. The anticipated mean cholesterol levels following treatments A and B are, respectively, 200 and 250 mg/dL, with a 20 mg/dL common standard deviation. To obtain a 95% confidence interval for the difference in cholesterol levels after treatment with (relative) precision $\varepsilon = 10\%$, the sample size should be at least $n = \frac{2 \times 1.96^2 \times 20^2}{0.10^2 (200 - 250)^2}$, that is 123 subjects per group.

1.4.2.3 Sample Size for Estimating a Regression-Based Reference Limit with a Given Precision

In some cases, the variable of interest might be indexed by a secondary variable. Suppose we have a continuum of distributions indexed by a covariate. For example, assume that we are studying BMI in a group of children of different ages. Instead of the mean BMI, we might be interested in other parameters of the BMI distribution, more particularly in the 95% reference limit of the BMI distribution for various ages. How many children should we sample in order to have a precise estimate of this reference limit and for every possible value of age? This question can be answered by applying linear regression techniques to estimate the reference limit as a function of age. Methods have been developed to estimate sample sizes for regression-based reference limits under various situations (Bellera and Hanley 2007). We provide an overview of these approaches, and refer the interested reader to this literature for additional details.

It is assumed that the mean value of the response variable of interest (e.g., BMI) varies linearly with the covariate (e.g., age), and that the response values are approximately normally distributed around this mean. The response variable and the covariate of interest are denoted by Y and X, respectively. Assume that at any given value x_0 of age, the mean value of interest, such as BMI, is an approximate linear function of X and that individual BMI values are normally distributed around this mean (the latter eventually after a suitable transformation) with constant variance: $Y \,|\, x_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\right)$. The 100$p$% reference limit for Y at this specific age point x_0 is given by: $Q_0 = \beta_0 + \beta_1 x_0 + z_p \sigma$, where z_p is the standard normal deviate corresponding to the 100p% reference limit of interest.

Given n selected individuals with data points $\left((x_i, y_i),\, i = 1, \ldots, n\right)$, a point estimator for the 100p% reference limit is given by: $\hat{Q}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + z_p s_{Y|X}$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by least-squares estimation of the regression coefficients β_0 and β_1, and $s_{Y|X}$ is the observed root mean square error.

Sample size estimation for regression-based reference limits requires defining the following parameters:

- The 100p% reference limit of interest, where $0 < p < 1$, and the corresponding one-sided standard normal deviate, z_p. For example, if we are interested in the 95% reference limit, the one-sided standard normal deviate is $z_{0.95} = 1.64$.

- The 100(1–α)% confidence interval for the reference limit of interest, and its corresponding two-sided standard normal deviate $z_{1-\frac{\alpha}{2}}$, where $0 < \alpha < 1$. For example, if we want the 95% confidence interval, then $\alpha = 0.05$ and $z_{1-\frac{\alpha}{2}} = 1.96$.
- The 100(1–β)% reference range, which encompasses 100(1–β)% of the values (e.g., BMI) as well as its corresponding two-sided standard normal deviate $z_{1-\frac{\beta}{2}}$, where $0 < \beta < 1$. For example, if we want the 95% reference range, then $\beta = 0.05$ and $z_{1-\frac{\beta}{2}} = 1.96$.
- The relative margin of error Δ, defined as the ratio of the width of the 100(1-α)% confidence interval for the reference limit to the width of the 100(1–β)% reference range. This means that we want a sample size large enough so that the width of the 100(1–α)% confidence interval for our reference limit is small when compared to the width of the 100(1–β)% reference range (we usually take $\alpha = \beta$).
- The design of the study, that is, the distribution of the covariate (e.g., age) in the sample investigated, which will influence the computation of the sample size.

Once the above parameters have been specified, we can estimate the sample size. Assume, first that we choose our sample so that the covariate (e.g., age) follows a uniform distribution. In such cases, the variance of the estimator $\hat{Q}$ of the reference limit is approximately equal to $\text{var}(\hat{Q}_0) = \frac{\sigma^2}{n}\left(4 + \frac{z_p^2}{2}\right)$. The width of the 100(1–α)% confidence interval for the 100p% reference limit at the *extreme* value of age is therefore $2z_{1-\frac{\alpha}{2}}\sigma\frac{\sqrt{4+\frac{z_p^2}{2}}}{\sqrt{n}}$. Assume we want a relative error of Δ, defined as the ratio of the width of the 100(1–α)% confidence interval for the reference limit to the width of the 100(1–β)% reference range. The width of the 100(1–α)% reference range is given by $2z_{1-\frac{\beta}{2}}\sigma$. Thus, if we want the ratio of the width of the 100(1–α)% confidence interval for the reference limit to the width of the 100(1–β)% reference range to be smaller than the relative error Δ, we require $\frac{z_{1-\frac{\alpha}{2}}\sqrt{4+\frac{z_p^2}{2}}}{z_{1-\frac{\beta}{2}}\sqrt{n}} \le \Delta$. That is, the minimum sample size, n, required to estimate the 100(1–α)% confidence interval for the 100p% reference limit, with a relative margin of error of Δ, when compared to the 100(1–β)% reference range, should be at least $n = \frac{z_{1-\frac{\alpha}{2}}^2\left(4+\frac{z_p^2}{2}\right)}{z_{1-\frac{\beta}{2}}^2\Delta^2}$. Similar formulae have been derived assuming other sampling strategies (Bellera and Hanley 2007). For example, instead of a uniform age distribution, one might take one-third of the sample at one age extreme, one-third at the midpoint, and one-third at the other age extreme. In this study design, the minimum

sample size requirement is then $\frac{z^2_{1-\frac{\alpha}{2}}\left(\frac{5}{2}+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}$. Similarly, we can also expect that the age distribution in the sample will follow a normal distribution. If we assume that the range of X is approximately 4 times the standard deviation of X, then we show that the sample size requirement becomes: $\frac{z^2_{1-\frac{\alpha}{2}}\left(5+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}$.

Notice that the previous formulae were derived under the "worst-case" scenario, that is, assuming that we are interested in estimating the reference limit at the extreme end of age, where the variability is highest, and thus the largest sample size is needed. If one is interested in the 100p% reference limit at the average age value (where the variability is minimized), then the sample size is reduced to $\frac{z^2_{1-\frac{\alpha}{2}}\left(1+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}$, for any given age distribution (or similarly, assuming a homogeneous population not indexed by a covariate). Put simply, information from either side of the average age adds strength to the information at the average age. In contrast, information at the extremes of the age distribution can only gather "strength" from one side of the age distribution; on the other side, there is infinite uncertainty.

Several other factors can impact the sample size (Cole 2006), such as for example the range of the covariate of interest. If one is interested in the height of children, a larger sample will be needed when considering birth to 18 years than when considering 5–12 years. Non-constant variability can also affect the sample size, since the variance of the estimator, $\text{var}(\hat{Q}_0)$, is proportional to $\frac{\sigma^2}{n}$, where σ^2 varies with age. This ratio can be made constant across age groups by ensuring that the sample size n is proportional to σ^2. Thus, at the ages at which the variability is increased, for example during puberty, the sample size needs to be increased appropriately to compensate (Goldstein 1986; Cole 2006). Similarly, in case of heteroscedasticity of the variable of interest across the covariate, regression techniques can be used to model the standard deviation as a function of the mean, and previous formulae can still be used as a rough guide for sample size planning. Notice that the variation σ^2 to be used in planning includes both the true inter-individual variability and the variability of the measuring instruments used: measurement tools with differing precisions will provide different sample size estimates. Finally, the nature of the relation between the covariate and the variable of interest can also affect the sample size, as more subjects will be needed to capture "wiggles" in the relation compared to a simple linear relation. If there is some nonlinearity in the covariate, such as for example a quadratic relationship, the formulae can also be accommodated by adjusting the point estimator of the reference limit of interest.

Example: We are interested in estimating a specific BMI reference limit as a linear function of age. Specifically, we wish to produce a 95% confidence interval $\left(z_{1-\frac{\alpha}{2}}=1.96\right)$ for the 95% BMI reference

limit ($z_{0.95} = 1.64$), with a relative margin of error $\Delta = 10\%$, when compared with the 95% reference range $\left(z_{1-\frac{\beta}{2}} = 1.96 \right)$. If age is uniformly distributed in the sample, then the minimum required sample size is $n = \frac{1.96^2 \left(4 + 1.64^2\right)}{1.96^2 0.10^2}$, i.e., we would need at least 536 observations to obtain this precise an estimate of the 95% BMI reference limit at any place in the age range. If, on the other hand, one is interested in the 95% reference limit only at the average age value, or in a homogeneous population not indexed by a covariate, then we would need at least $n = \frac{1.96^2 \left(1 + 1.64^2\right)}{1.96^2 0.10^2}$, that is, 236 observations.

1.5 Principles of Hypothesis Testing

We have presented formulae for the estimation of the sample size required to estimate various parameters with a desired degree of precision. Similarly, one may want to ensure that a sufficient number of subjects are available to show a difference between two parameters. We discuss here some basic principles of hypothesis testing which can be found in introductory statistical textbooks, as well as works devoted to clinical trials or epidemiology (Lachin 1981; Friedman et al. 1998; Armitage et al. 2002).

One wishes to compare two groups, A and B, with true prevalence rates, π_A and π_B. From a statistical point of view, this comparison is expressed in terms of a **null hypothesis**, called H_0, which states that no difference exists between the two groups: H_0: $\pi_A - \pi_B = 0$. Hypothesis testing consist in testing whether or not H_0 is true, more specifically, whether or not it should be rejected. Thus, until otherwise proven, H_0 is considered to be true. The true prevalence rates, π_A and π_B are unknown. If two groups of subjects are properly sampled, one can obtain appropriate estimates P_A and P_B. Although π_A and π_B might not differ, it is possible that by chance alone, the observed proportions P_A and P_B are different. In such cases, one might falsely conclude that the two groups have different prevalence rates. Such a false-positive error is called **a type 1 error**, and the probability of making such an error corresponds to the **significance level** and is denoted by α. The probability of making a type 1 error should be minimized. However, decreasing the significance level increases the sample size. The probability of observing a difference as extreme as or more extreme than the difference actually observed, given that the null hypothesis is true, is called the ***p-value*** and is denoted by P. The null hypothesis H_0 will be rejected if $p < \alpha$.

If the null hypothesis is not correct, then an **alternative hypothesis**, denoted by H_A must be true, that is H_A:$\pi_A - \pi_B = \delta$, where $\delta \neq 0$. It is possible that by chance alone, the observed proportions P_A and P_B differ only by a small amount. As a result, the investigator may fail to reject the null hypothesis. This false-negative error is called **a type 2 error**, and the probability of making such error is denoted by β. The probability of correctly accepting H_0 is thus $1-\beta$ and is referred as the **power**. It defines the capability of a statistical test to reveal a given difference between two parameters, if this difference really exists. The power depends on the size of the difference we wish to detect, the type 1 error, and the number of subjects. That is, if the type 1 error and the sample size are held constant, a study will have a larger power if one wishes to detect a large difference compared to a small one.

1.6 Power-Based Sample Size Estimation

The number of subjects needed should be planned carefully in order to have sufficient power to detect significant differences between the groups considered. To provide a power-based (or test-based) estimation of sample size requires defining a priori the difference one wishes to detect, the desired significance level and the desired power of the test.

As will be discussed below, sample size formulae involve the ratio of the variance of the observations over the difference one wishes to detect, or more generically, a noise–signal ratio, where the signal corresponds to the difference one wishes to detect, and the noise (or uncertainty) is the sum of all the factors (sources of variation) that can affect the signal (Sackett 2001).

Note that we only discuss the case of independent observations. In clinical trials, for example, it may not always be possible to randomize individuals. For example, a physical activity intervention might be implemented by randomizing schools. Individuals are then grouped or clustered within schools. They cannot be considered as statistically independent, and the sample size needs to be adapted since standard formulae underestimate the total number of subjects (Donner et al. 1981; Friedman et al. 1998).

1.6.1 Dichotomous Variables

1.6.1.1 Sample Size for Comparing a Proportion to a Theoretical Value

The variable of interest is dichotomous. For example, one is interested in the prevalence of obese children π_1. Based on a random sample, the objective is to compare this proportion to a target (fixed) value p_0. The null and alternative hypotheses are given respectively by $H_0: \pi_1 = p_0$ and $H_A: \pi_1 \neq p_0$. The minimum number of subjects needed to perform this comparison assuming an observed prevalence rate p_1, a significance level α and power $1-\beta$ is $n = \dfrac{\left(z_{1-\frac{\alpha}{2}}\sqrt{p_0(1-p_0)} + z_{1-\beta}\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}$.

Example: One is interested in evaluating a specific diet aimed at reducing weight in obese subjects. A success is defined as reducing a BMI down to 25 or lower. One wishes to test whether this new diet has a better success rate than the standard diet for which the efficacy rate is known to be $p_0 = 20\%$. The anticipated efficacy rate of the new diet is $p_1 = 40\%$. The sample size needed to show a difference between the efficacy rate of the new diet and the target efficacy rate p_0 assuming a significance level $\alpha = 0.05$ $\left(z_{1-\frac{\alpha}{2}} = 1.96\right)$, and power $1-\beta = 0.90$ ($z_{1-\beta} = 1.28$) is at least $n = \dfrac{\left(1.96\sqrt{0.20(1-0.20)} + 1.28\sqrt{0.40(1-0.40)}\right)^2}{(0.40-0.20)^2}$, i.e., 36 patients. If the anticipated efficacy rate is $p_1 = 30\%$, that is, one wishes to detect a smaller difference, then the sample size must be increased to at least $n = \dfrac{\left(1.96\sqrt{0.20(1-0.20)} + 1.28\sqrt{0.30(1-0.30)}\right)^2}{(0.30-0.20)^2}$, that is, 137 subjects.

Note that the calculated sample sizes are quite low. This is because we are comparing one sample to one historical (or literature-based) sample. In practice, one will usually be comparing two samples (next section), as in clinical trials. In this case, the resulting sample size is much higher as variability of the second sample has to be accounted for.

1.6.1.2 Sample Size for Comparing Two Proportions

The outcome of interest is dichotomous and two independent groups of equal size are being sampled and compared. The objective is to detect a difference between two proportions π_0 and π_1, that is, the null and alternative hypotheses are given respectively by H_0: $\pi_1-\pi_0=0$ and H_A: $\pi_1-\pi_0=\delta$, where $\delta\neq 0$. The size of each sample required to detect an anticipated difference p_1-p_0, assuming a significance level α, and power β is $n=\dfrac{\left(z_{1-\frac{\alpha}{2}}\sqrt{2\overline{p}(1-\overline{p})}+z_{1-\beta}\sqrt{p_0(1-p_0)+p_1(1-p_1)}\right)^2}{\left(p_1-p_0\right)^2}$, where $\overline{p}$ is the average of the two anticipated proportions p_0 and p_1. Similar formulae are available for the case of unequal sample sizes (Piantadosi 2005).

Example: A trial is set up to compare two physical activity interventions with anticipated success rates of 50% and 30%. The sample size per group needed to show a difference between the two interventions assuming a difference in success rates $\delta=50\%-30\%=20\%$, a significance level $\alpha=0.05$ $\left(z_{1-\frac{\alpha}{2}}=1.96\right)$, and power $1-\beta=0.90$ $\left(z_{1-\beta}=1.28\right)$ is $n=\dfrac{\left(1.96\sqrt{0.40(1-0.40)}+1.28\sqrt{0.30(1-0.30)+0.50(1-0.50)}\right)^2}{0.20^2}$ that is, 84 subjects per group.

1.6.2 Continuous Variables

1.6.2.1 Sample Size for Comparing a Mean to a Theoretical Value

The variable of interest is continuous. For example, one is interested in the mean waist circumference, μ_1, following a physical activity intervention. Based on a random sample of subjects, the objective is to compare this mean circumference to a target mean value m_0. The null and alternative hypotheses are thus given respectively by H_0: $\mu_1=m_0$ and H_A: $\mu_1\neq m_0$. For this one sample problem, the number of subjects needed to perform this comparison assuming an anticipated mean value m_1, a standard deviation σ, significance level α and power $1-\beta$ is $n=\dfrac{\sigma^2\left(z_{1-\frac{\alpha}{2}}+z_{1-\beta}\right)^2}{\left(m_1-m_0\right)^2}$. Unless the variance σ^2 is known, one can use a literature-based or experience-based s^2 as an estimate of the population variance.

Example: One wishes to test whether mean waist circumference in a given population following a new physical activity intervention is reduced compared to a standard intervention. Following the standard intervention, the mean waist circumference is known to be about $m_0 = 140$ cm. It is anticipated that the new intervention will reduce this circumference to $m_1 = 130$ cm. The sample size needed to show a difference between the mean waist circumference with the new intervention and a target value m_0, assuming a significance level $\alpha = 0.05$ $\left(z_{1-\frac{\alpha}{2}} = 1.96\right)$, a 90% power $\left(z_{1-\beta} = 1.28\right)$ and standard deviation SD = 20cm, is $n = \frac{20^2\left(1.96+1.28\right)^2}{\left(130-140\right)^2}$, that is, 42 patients.

1.6.2.2 Sample Size for Comparing Two Means

The variable of interest is continuous and two independent groups of equal size are being sampled and compared. The objective is to detect a difference between two means, that is, the null and alternative hypotheses are given respectively by $H_0: \mu_1 - \mu_0 = 0$ and $H_A: \mu_1 - \mu_0 = \delta$, where $\delta \neq 0$. The sample size of each group required to detect this difference assuming an anticipated difference $m_1 - m_0$, common variance σ^2, significance level α and power β is $n = \frac{2\sigma^2 \times \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{(m_1 - m_0)^2}$. Unless the variance σ^2 is known, one can use a literature-based or experience-based s^2 as an estimate of the population variance.

Example: In their example, Armitage et al. are interested in comparing two groups of men using the forced expiratory volume (FEV) (Armitage et al. 2002). From previous work, the standard deviation of FEV is 0.5 L. A two-sided significance level of 0.05 $\left(z_{1-\frac{\alpha}{2}} = 1.96\right)$ is to be used with an 80% power ($z_{1-\beta} = 0.842$). In order to show a mean difference of 0.25 L between the groups, and assuming samples of equal sizes, the total number of men should be at least $n = \frac{2 \times 0.5^2 \times \left(1.96 + 0.842\right)^2}{0.25^2}$, that is, 63 men per group.

1.7 Other Parameters, Other Settings

Sample sizes can be estimated for various parameters and under various settings. As such, it is not possible to cover all possible situations into a single book chapter! We have reviewed formulae for the estimation of sample sizes for commonly used parameters such as means, proportions and reference limits. Other parameters such as time-to-event outcomes (Freedman 1982; Schoenfeld 1983; Dixon and Simon 1988), correlation coefficients (Bonett 2002), concordance coefficients (Donner 1998), or even multiple endpoints can be considered (Gong et al. 2000). Similarly, methods for calculating sample size assuming other designs have been investigated. Instead of detecting a specific difference, one might be interested in showing equivalence or noninferiority (Fleming 2008); observations may be clustered (Donner et al. 1981; Hsieh 1988), etc. Sample size estimation procedures have been developed for these settings and we refer the interested reader to specialized literature or general works on sample size (Friedman et al. 1998; Machin et al. 1997; Altman et al. 2000; Piantadosi 2005).

Finally, although they should be used with caution, several statistical software packages, such as nQuery® (nQuery Advisor ® 6.0), or East® (Cytel), are available to compute sample size and power for means, proportions, survival analysis, etc.

1.8 Application to Other Areas of Health and Disease

The methods presented in this chapter can be applied to many areas of research and study design, including anthropometry, but also fundamental biology, epidemiology, clinical trials, social sciences, demography, or economics.

Summary Points

- Before estimating a sample size, the nature and the distribution of the variable of interest must be defined. The parameter of interest (mean, proportion, reference limit) can then be identified.
- Before estimating a sample size, the type of sample has to be identified: one single sample? Two samples?
- Sample size estimation can be either precision-based or power-based.
- When performing precision-based sample size estimation, the anticipated value of the parameter as well as the level of significance and the precision (absolute or relative) must be defined a priori.
- When performing power-based sample size estimation, the anticipated value of the parameter as well as the level of significance and the power must be defined a priori.

Key Features of Sample Size Estimation

Table 1.3 Sample-size required for the estimation of a 100(1–α)% confidence interval for various parameters (*assuming an absolute margin of error* ME)

Parameter of interest	Sample size
Proportion (assuming an anticipated value P)	$n = \frac{z^2_{1-\frac{\alpha}{2}} p(1-p)}{\mathrm{ME}^2}$
Difference between 2 proportions (assuming independent samples with the same sample size and anticipated proportions p_1 and p_2)	$n\ (\text{per group}) = \frac{z^2_{1-\frac{\alpha}{2}} \left(p_1(1-p_1) + p_2(1-p_2)\right)}{\mathrm{ME}^2}$
Mean (assuming sample variance s^2)	$n = \frac{z^2_{1-\frac{\alpha}{2}} s^2}{\mathrm{ME}^2}$
Difference between 2 means (assuming independent samples with the same sample size and common sample variance s^2)	$n\ (\text{per group}) = \frac{2\, z^2_{1-\frac{\alpha}{2}} s^2}{\mathrm{ME}^2}$
Regression-based reference limit for the 100p% reference limit (assuming a uniform distribution for the covariate and a relative margin of error of Δ when compared to the 100(1–β)% reference range)	$n = \frac{z^2_{1-\frac{\alpha}{2}} \left(4 + \frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}} \Delta^2}$

Table 1.4 Sample-size required for the comparison, via a test, of various parameters (assuming significance level α, and power $1-\beta$

Comparison of interest	Sample size
Single proportion π_1 (with anticipated value p_1) to a theoretical value p_0	$n = \dfrac{\left(z_{1-\frac{\alpha}{2}}\sqrt{p_0(1-p_0)} + z_{1-\beta}\sqrt{p_1(1-p_1)}\right)^2}{(p_1-p_0)^2}$
Two proportions (assuming independent samples of equal size, with anticipated proportions p_0 and p_1 and where $\overline{p} = (p_0+p_1)/2)$	n (per group) $= \dfrac{\left(z_{1-\frac{\alpha}{2}}\sqrt{2\overline{p}(1-\overline{p})} + z_{1-\beta}\sqrt{p_0(1-p_0)+p_1(1-p_1)}\right)^2}{(p_1-p_0)}$
Single proportion μ_1 (with anticipated value m_1) to a theoretical value m_0 (assuming sample variance s^2)	$n = \dfrac{s^2\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{(m_1-m_0)^2}$
Two means (assuming independent samples of equal size with common sample variance s^2)	n (per group) $= \dfrac{2s^2 \times \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{(m_1-m_0)^2}$

References

Altman DG, Machin D, Bryant T, Gardner S. Statistics with confidence: Confidence Intervals and statistical Guidelines. 2nd ed. BMJ Books; 2000.

Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4 ed. Blackwell Science; 2002.

Bellera CA, Hanley JA. A method is presented to plan the required sample size when estimating regression-based reference limits. J Clin Epidemiol. 2007;60:610–5.

Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21:1331–5.

Cole TJ. The International Growth Standard for Preadolescent and Adolescent Children: Statistical considerations. Food Nutr Bull. 2006;27:S237–3.

Dixon DO, Simon R. Sample size considerations for studies comparing survival curves using historical controls. J Clin Epidemiol 1988;41:1209–13.

Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. Stat Med. 1998;17:1157–68.

Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. Am J Epidemiol. 1981;114:906–14.

Elveback LR, Guillier CL, Keating FR. Health, normality, and the ghost of Gauss. JAMA. 1970;211:69–75.

Fleming TR. Current issues in non-inferiority trials. Stat Med. 2008;27:317–32.

Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med. 1982;1:121–9.

Friedman L, Furberg C, DeMets DL. Fundamentals of clinical trials. 3rd ed. New York: Springer-Verlag; 1998.

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;292:746–50.

Gardner MJ, Altman DG. Estimating with confidence. BMJ. 1988;296:1210–1.

Goldstein H. Sampling for growth studies. In: Falkner F, Tanner JM, eds. Human growth: a comprehensive treatise, 2nd ed. New-York: Plenum Press. 1986. p. 59–78.

Gong J, Pinheiro JC, DeMets DL. Estimating significance level and power comparisons for testing multiple endpoints in clinical trials. Control Clin Trials. 2000;21:313–29.

Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine, Vol 146 of Statistics: textbooks and Monographs. New York: Marcel Dekker; 1995.

Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. Stat Med. 1988;7:1195–201.

Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials. 1981;2:93–113.

Last J. A Dictionnary of Epidemiology. 4th ed. Oxford University Press; 2001.

Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies. 2nd ed. London: Blackwell Science; 1997.

Piantadosi S. Clinical Trials: A Methodologic Perspective. 2nd ed. Hoboken, New Jersey: John Wiley and Sons; 2005.

Sackett DL. Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). Canadian Medical Association Journal 2001;165:1226–36.

Schoenfeld D. Sample-Size Formula for the Proportional-Hazards Regression Model. Biometrics. 1983;39:499–503.

World Health Organization. Department of nutrition for health and development. Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for age: Methods and development. WHO Press; 2006.

Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Stat Methods Med Res. 1999;8:93–112.