Clinical Research Methods

Sampling techniques, confidence intervals, and sample size

ABHAYA INDRAYAN, PIYUSH GUPTA

INTRODUCTION

The concept of sampling is a familiar one. A cook examines a few grains of rice to find out whether or not they are properly cooked. The aim of sampling is to choose a relatively small-sized sample that adequately represents the entire population. The results obtained can then be extrapolated to the entire population that it represents. 'Population' here refers to the target group to which the results of the investigation are intended to be applied.

ADVANTAGES AND LIMITATIONS OF SAMPLING

Sampling can be highly advantageous due to its inherent feasibility and cost-effectiveness. Results can be obtained fast and tend to be reliable, because technologically superior methods and close monitoring can be employed on a relatively smaller number of subjects in a sample than in a large population. However, sampling has certain limitations. At times, the sample may not truly represent the population. It may create a sense of discrimination. Sampling may not be feasible in a small population and is not necessary where a complete count of the population is needed.

Since individuals in the sample differ from one another, it is natural that the samples also vary from one another. *Sampling error* refers to the fluctuation of results amongst different samples in the same population designed to measure the same parameter. This error is not a mistake but signifies fluctuation only.

Example 1: If a sample of 300 healthy men 60–64 years of age from a population have an average systolic blood pressure (BP) of 142 mmHg, it is possible that another sample of 300 from the same population yields an average of 139 mmHg. But how likely is it that the average in the new sample is as low as 128 mmHg or as high as 155 mmHg?

The magnitude of this 'error' depends primarily on three factors:

1. *The method of sampling.* The subjects should be selected in a manner that a wide spectrum gets adequate representation. Then repeated samples may give nearly the same picture.
2. *Variability between the subjects in the population.* If cholesterol levels differ widely from person-to-person, then obviously the samples too would reflect the same variability.
3. *The size of the sample.* When the sample includes a large number of subjects, the picture obtained from one sample is not likely to vary much from another sample of the same size, because both tend to be fair representations of the population. This cannot be said for small samples.

University College of Medical Sciences, Dilshad Garden, Delhi 110095, India
ABHAYA INDRAYAN Department of Biostatistics and Medical Informatics
PIYUSH GUPTA Department of Paediatrics
Correspondence to ABHAYA INDRAYAN

TERMINOLOGY

Sampling fraction

The number of subjects in the sample is denoted by n and the number of subjects in the entire target population by N. The ratio n/N is called the sampling fraction.

Unit

This may be related to the sampling or inquiry. The *sampling unit* is the one used for selection of subjects. The *unit of inquiry* is the subject on which information is obtained. For example, in a community survey on protein–energy malnutrition, the sampling unit could be a family but the unit of inquiry may be a child <5 years old. One sampling unit can have multiple or no units for inquiry.

Sampling frame

This refers to the list of all sampling units contained in the target population. The units must be mutually exclusive and the frame should be an exhaustive list. If there is a study on hypertensives and diabetics, a person who is both cannot be listed twice. The list may separately include those who have both the diseases, either of the diseases and none of the diseases. Inclusion and exclusion criteria must be fully known to the sampler. Preparation of a frame requires precise definition of the unit as well as the population.

SAMPLING METHODS

The literature on sampling methods is extensive.[1] Sampling can either be random or purposive.

Purposive sample

A sample of volunteers, as in a Phase I clinical trial, is a purposive sample. Such a trial may help to identify the efficacy and adverse effects of a certain intervention. However, the information obtained could be biased and the results obtained from such samples may not be amenable to generalization. For valid extrapolation, it is necessary that a random sample is chosen.

Random or probability sample

The inclusion of a particular unit in the sample cannot be predicted in this sampling and would depend, at least to some extent, on chance. Random sampling does not ensure that the characteristics of the sample units would coincide exactly with those in the population but it ensures a known probability that their divergence lies within a given limit.

SIMPLE RANDOM SAMPLING

When the scheme is such that each unit of the population has the same chance of being included in the sample, it is called simple random sampling (SRS). SRS is like picking n chits from a lot

containing N chits, numbered 1 through N. A more scientific method is to use random numbers. Random numbers can be easily generated on a computer, and are also available in table form in standard textbooks.

A prerequisite for SRS is the availability of the sampling frame. Preparation of this in many cases can be an expensive exercise. If a study involves several hospitals, each hospital will have a list of its own patients but a joint list of all the patients in all the hospitals may not be available at one place. The next problem is that the selected units can be physically far apart—in different areas, admitted to different hospitals or attended to in clinics in different locations. Thirdly, there is no guarantee that an SRS will adequately represent different segments of the target population. To overcome this problem, stratified sampling is used.

STRATIFIED RANDOM SAMPLING

A drawback of SRS is that it may fail to adequately represent one or more subgroups of interest. For example, in a study on relationship of maternal complications with parity, it is necessary that women of different parity are included in the sample. The definition of the sampling unit in this case could be a currently pregnant woman reporting to a particular group of antenatal clinics. In this case, an SRS of 60 can yield a sample in which women with parity 4 remain under-represented or are omitted altogether. Therefore, the procedure should be to first divide the frame by parity status such as 1, 2, 3, 4, 5 and 6+ and then draw an independent SRS of size 10 from each division. Such a division of the frame is called *stratification* and each division a *stratum.* The researcher decides how many units are to be selected from different strata; they need not be equal. This sampling procedure is called stratified random sampling (StRS).

The characteristic chosen for stratification is either the one suspected to affect the variable under study, or that which makes groups of interest for which different results are required. After this, the sample would adequately represent the stratifying characteristic but not necessarily other factors that may be of consequence. For example, in a study on diabetes mellitus, levels of plasma glucose may stratify the subjects, but the sample may still not be representative for obesity, age–gender, co-existing diseases, patient cooperation, etc. All these can affect the prognosis or the outcome. Thus, due care is always needed in extrapolation of results even after adopting stratified sampling.

MULTI-STAGE RANDOM SAMPLING

In studies that involve large populations, it is sometimes helpful to draw the sample in stages. If the target population consists of subjects spread all over a state, a small number of districts are selected in the first stage. In the next stage, some blocks, colonies or hospitals can be selected from the chosen districts; and finally, the subjects are identified from the selected blocks/colonies/hospitals. Thus, there are sampling units of various sizes. When sampling is done in stages from bigger to smaller units within the units selected at a previous stage, it is called multi-stage sampling. When the selection in each stage is at random, this becomes multi-stage random sampling (MRS).

Example 2. In a study to find out the prevalence of smoking in women 20 years or older in a particular state with, say, a million families, we may, for example, first select four districts by the random method, then 40 census blocks within each selected district and 10 families within each of the selected blocks. All women 20 years or more in the selected families could be the unit of inquiry, though the sampling units are districts, blocks and families. Some families may have two or more units of inquiry and some none at all; most may have just one. If there are many families with two or more eligible women then this can produce a clustering effect. This is discussed in the next section.

In the above example, a total of 4×40×10=1600 families could be in the sample. This may seem to be an extremely small number compared to a total of a million families in the state, yet could provide a fairly precise estimate of the prevalence of smoking among women in the state.

SRS v. MRS

If an SRS of 1600 families is chosen out of a million, the selected families may be scattered all over the state, say, in 200 census blocks. A block may have to be visited for just one family. This could mean a substantially higher cost of travel as well as loss of time. However, in the case of MRS, only 4 districts require to be visited and the survey workers will concentrate in 40 blocks within each district. Thus, the major advantage of MRS is the reduced cost and time saved due to less travel. The second advantage is that a full sampling frame of the smaller units is not needed. In our example, the frame required is the list of all districts in the state, the list of blocks in the selected districts, and the list of families in the selected blocks only. However, in the case of SRS, the frame will be the list of all families in the entire state. Preparation of such a frame could be a major exercise in some situations. The third advantage is that, in most situations, a smaller sample chosen by MRS may be sufficient to achieve a good precision, relative to SRS.

CLUSTER RANDOM SAMPLING

When the primary sampling units are not large, i.e. when they generally contain a small number of subjects, then it is sometimes advisable that these units are not sampled further. All the elements in the selected primary units are then surveyed. This tends to increase the total number of subjects in the sample without a corresponding increase in cost. When this is done, it is convenient to understand a primary unit as a cluster. If a population comprises a total of N clusters, then n clusters out of N are randomly selected. If the ith cluster has M_i subjects, then a total of $\sum_i M_i$ elements of these n clusters are investigated. This is called cluster random sampling (CRS).

Besides increasing the sample size and thus the precision, without a corresponding increase in cost, the other advantage relative to SRS is that the sampling frame of the elements is not required. The only frame required is the list of clusters. Since a survey of subjects within a cluster is quick, CRS is sometimes considered to be a rapid assessment method. The World Health Organization (WHO) recommends this kind of sampling to estimate the percentage of children immunized in a community, particularly in developing countries. Their recommended strategy is 30 clusters of 7 each, also called 30×7 sampling. Bennett *et al.*[2] have given some useful details of this method.

Clustering effect

A major disadvantage of CRS is that the elements within a cluster tend to be similar to one another and produce a clustering effect. This effect reduces the chances of getting the full spectrum of subjects in the sample. To compensate for this, a larger sample may be required relative to SRS. However, it sometimes happens that even a large sample chosen by CRS is less expensive to investigate than a small sample chosen by SRS. An inter-cluster comparison is not valid in the case of CRS.

In CRS, it is customary to divide the population into clusters of approximately equal size. However, this is not a prerequisite.

Example 3. **For a survey on prevalence of poor vision (visual acuity < 6/36 in the better eye with corrective glasses, if any) in persons > 50 years of age in a district with half a million population, 20 clusters of size 30 each are selected as follows:**

1. **A list of census blocks is prepared along with the population of each, which is cumulatively added.**
2. **Since the sampling fraction is 1 cluster per 25 000 population, one number less than or equal to 25 000 is randomly selected. Then 25 000 are sequentially added every time in a systematic fashion and thus a sample of 20 numbers is obtained. Twenty blocks containing the chosen 20 numbers are selected from the list made in step (1). These blocks are now in the sample.**
3. **A start is made for home visits from a geographically random point in each of the selected blocks and the first 30 persons more than 50 years of age residing in contiguous houses are listed and examined for visual acuity.**

Note the following features of this CRS:

1. The frame required is only the list of census blocks, which in any case is generally available. No listing of households or of persons >50 years of age is required.
2. The selection of blocks is made on the basis of the size of the population. This is inherent in step (2) of the above example. Blocks with a bigger population have a higher chance of being included in the sample. This is called sampling with *probability proportional to size (pps)*. The size in this case is the population while the subjects are persons >50 years of age. It is reasonable to expect that this age group would have nearly the same proportion in each census block. The pps sampling makes the estimate self-weighted and the usual sample proportion becomes a statistically valid estimate of the population proportion.
3. Starting from a geographical random point ensures random sampling but may require a map of each of the selected blocks. A map of cluster boundaries is not required in this example.
4. The houses to be visited are contiguous. This should make the survey fast.

SYSTEMATIC RANDOM SAMPLING

In the previous example on CRS, the first unit was randomly selected and the others automatically included on the basis of the *sampling interval* $I=N/n$. Such a scheme is called systematic random sampling (SyRS). The units are numbered I to N in any order. If I is not an integer then the integer part is taken. The first unit is selected at random from the first I units. Suppose this is the rth unit. Then the subsequent units in the sample are $(r+I)$th, $(r+2I)$th, etc. Thus the first selected unit determines the entire sample.

Example 4. **If there are $N=101$ subjects in the target population out of which $n=8$ are proposed to be selected by SyRS, the sampling interval is 101/8 and its integer part is 12. If the randomly selected unit (subjects in this case) of the first 12 is the 9th then the remaining units are 21, 33, 45, 57, 69, 81 and 93.**

While this method works well when the sampling interval is an integer, two types of anomalies can occur in other cases:

1. The last few units may never get a chance to be included in the sample. In the above example, if the first selected number is the maximum 12 then the last number in the sample is 96. Thus the unit numbers 97 to 101 can never be selected.
2. In some cases, the sample size may be exceeded. In this example, if the first randomly selected unit is 4 then the last would be 100. The sample would then contain 9 units instead of the stipulated 8. However, this demerit is not serious when n is large, say more than 30. A way out is *circular sampling* in which a start is made from a random number between I and N, and every Ith thereafter is selected in a cyclical manner, i.e. $(N+I)$th unit is the first unit again. This is done until n units are selected.

Merits and demerits

SyRS is easy and speedier to execute and does not need a full sampling frame. In some cases, the method can yield a much more representative sample than SRS. The chosen units are equally spaced till the end and every section of the population is likely to get represented. On the negative side, the SyRS could yield a biased sample if a periodicity or trend is hidden in the subjects as one moves from I to N. As in the case of SRS, this method too can fail to adequately represent some groups of interest.

- For a rare condition (e.g. cancer of the colon), it may be advisable to continue sampling till such time that a predetermined adequate number of cases are available for investigation. This procedure is called *inverse sampling*. The total sample in this case can become exceedingly large and cannot be fixed in advance. In this kind of sampling, the estimate and its variance change. A brief description is given by Som.[3]
- Selection of subjects is sometimes made from readily available groups such as hospital patients. The findings based on such subjects cannot be applied to the general population because hospital patients are highly 'selected'. Any such sample will have a selection bias.
- The sampling frame should be complete and up-to-date for sampling to be successful.
- While small studies may be based on subjects selected by one of the above methods, medium and large studies would generally be based on a mix of two or more methods.

INTERSAMPLE VARIABILITY

In an empirical science such as medicine, a decision is made on the basis of the samples, despite the presence of sampling fluctuations, that too mostly on the basis of just one sample. Fortunately, statistical methods allow us to draw valid conclusions by assessing the expected magnitude of intersample variability on the basis of just one sample. This is done by calculating the standard error (SE) of the desired statistic as explained below.

Depending upon whether the variable is qualitative or quantitative, the parameter generally of interest is proportion π (or probability) or mean μ, respectively. In case of nutritional anaemia, the interest could either be in proportion π (or percentage) of subjects responding to iron–folic acid supplementation, or in mean rise μ in haemoglobin (Hb) or haematocrit (Hct) levels. If in a sample of 200 subjects, the proportion responding is $p=0.30$ (or 30%) or the mean rise in Hb level is $\bar{x}=0.8$ g/dl after supplementation for 100 days, then it is generally believed that the entire target population would also show a similar picture. The sample values p and $\bar{x}$ are considered to be *point estimates* of π and μ, respectively.

Sampling fluctuation can be studied by analysing the behaviour of p and $\bar{x}$ from sample-to-sample.

Standard error of p and $\bar{x}$

Standard error (SE) is the measure of variability of estimates of p and $\bar{x}$ from sample-to-sample. This special term helps to distinguish inter-individual variability (measured by standard deviation) from intersample variability (measured by SE). The larger the SE, the more is the variability and less the confidence in the sample results.

The SE is calculated on the basis of all possible samples of specific size n from the identified target population. However, these samples are not actually drawn. Statistical theory helps to obtain SE on the basis of just one sample, provided it is randomly drawn at least at one stage. This underscores the need to work with random samples. In case of SRS, the SEs are estimated as follows:

$$\text{SE}(p) = \sqrt{\frac{p(1-p)}{n}} \qquad (1)$$

and

$$\text{SE}(\bar{x}) = \frac{s}{\sqrt{n}} \qquad (2)$$

where s is the sample SD and computed with the denominator $(n-1)$. This adjustment in the denominator helps to achieve a more accurate estimate in the long run (called 'unbiased' in statistical parlance). The following comments in this context are useful.

1. Estimated SE (p) is maximum when $p=0.5$, and smaller when p is either small or large. However, its interpretation requires additional care.

   For $n=100$ and $p=0.05$, SE $(p)=\sqrt{(0.05\times0.95)/100}=0.022$

   For $n=100$ but $p=0.25$, SE $(p)=\sqrt{(0.25\times0.75)/100}=0.043$.

   In absolute terms, the SE in the first case is nearly half of what it is in the second case. In relative terms though, the first SE is 44% of p while the second is only 17% of p. Thus, the first is higher in a relative sense. This has serious repercussions on confidence intervals.
2. The estimate (1) fails when $p=0$ because then SE=0. This is discussed later.
3. SE $(\bar{x})$ is large when σ (SD of individual measurement in a population) is large. That is, if the individuals vary too much from one another then the sample means too would exhibit a large variation from sample-to-sample. Conversely, if the individual measurements are nearly alike or homogeneous, then the mean of two samples would be nearly similar.
4. Both the SEs are inversely proportional to the square root of the sample size n. They decrease as n increases. This means that two samples containing 100 subjects each from the same population would not differ as much as two samples containing only 20 subjects each. A larger sample size thus increases the confidence in the sample results. However, this increase could be counter-productive in terms of escalation of cost. Thus, a balanced approach is needed.

Sampling distribution of p and $\bar{x}$

As mentioned, the values of proportion p and mean $\bar{x}$ tend to vary from sample-to-sample. Just as individual measurements have their distribution pattern such as Gaussian, skewed or U-shaped, the sample proportion and sample mean too have a specific distribution pattern when many samples are available. This is called the sampling distribution. It has been theoretically established that in almost all situations in medical practice, this pattern tends to become Gaussian when n is large even when the underlying distribution among individuals is not Gaussian. This statistical result is called the *central limit theorem (CLT)* and is very useful in drawing an inference when n is large. The following criteria help in deciding whether n is large or not.

For proportion, n is large if

$$np\geq8 \text{ and } n(1-p)\geq8, \text{ when } p\geq0.01 \text{ and } (1-p)\geq0.01 \qquad (3)$$

For mean, n is large if

$$n\geq30 \qquad (4)$$

The criterion (3) could mean a very large n if p is really small. If $p=0.3$, then n must be a minimum of 26 and if $p=0.02$ then n must be at least 400. If $p<0.01$, then an approximation called Poisson[4] is used for a large n because Gaussian approximation is not sufficiently good for such a small p unless n is very large.

The situation is not so simple for a small n. The sampling distribution of p is based on a binomial distribution. This can be treated as Gaussian for a large n but not for a small n. The underlying distribution in this case is called Bernoulli. The sampling distribution of $\bar{x}$ also would be non-Gaussian for a small n when the underlying distribution is non-Gaussian. Different sets of methods called non-parametric methods are generally required in this case. The sampling distribution of $\bar{x}$ would be Gaussian even for small n, if the underlying distribution were Gaussian.

CONFIDENCE INTERVALS

When an average value or a proportion (or any other quantity such as ratio) is calculated from a random sample, the range within which the corresponding population parameter is expected to lie with a given probability in repeated samples can be estimated. This probability is called confidence and the range so obtained a confidence interval (CI). This is also known as *interval estimate.*

A CI gives a range with a hope that it will include the parameter of interest. The confidence level associated with the interval (say 90%, 95% or 99%) gives the percentage of all such possible intervals that will actually include the true value of the parameter. A CI tells us what to expect in the long run. It does not say anything about a particular sample.

Confidence interval for π, μ and their differences (large n)

CI for proportion π: The 95% CI for a large n can almost invariably be obtained as estimate ±2SE. In the case of π, this becomes

$$\left(p-2\sqrt{\frac{p(1-p)}{n}},\ p+2\sqrt{\frac{p(1-p)}{n}}\right) \qquad (5)$$

This is valid only when the observed sample proportion p is neither zero nor one. It also requires that p and $(1-p)$ are both at least 0.01.

Example 5. **The management of cases of bronchiolitis in infants may become easier if the course of the disease could be predicted on the basis of the condition at the time of hospital admission. One simple criterion for this could be the respiratory rate (RR). Suppose in a random block of 80 consecutive cases of bronchiolitis coming to a hospital with RR≥68, a total of 51 (64%) are ultimately observed to have a serious form of the disease, i.e. they either had a prolonged hospital stay, developed some complication, required endotracheal intubation or mechanical ventilation, or died. The percentage observed in this sample is 64. Other samples from the same hospital or from the same area may give a different percentage. What could be the percentage of cases with a serious form of disease in the entire 'population' of patients admitted to the hospital with a diagnosis of bronchiolitis and**

RR≥68 per minute? Since $n=80$ and $p=0.64$, np and $n(1-p)$ are large enough to ensure a Gaussian pattern. The ± 2SE limits in this example are

$$p-2\text{SE}(p)=0.64-2\sqrt{0.64(0.36)/80}=0.53 \text{ and}$$

$$p+2\text{SE}(p)=0.64+2\sqrt{0.64(0.36)/80}=0.75$$

Thus, the confidence interval is 0.53, 0.75. This implies that there is between 53% and 75% chance that a case of bronchiolitis with RR≥68 per minute at the time of hospitalization will require special handling. Suppose that 6% of those with RR≥68 per minute in the above example fail to survive. The 95% CI for the proportion dying is

$$(0.06-2\sqrt{(0.06\times 0.94)/80},\ 0.06+2\sqrt{(0.06\times 0.94)/80})$$

or (0.01, 0.11)

Thus, the actual fatality rate could be anywhere between 1% and 11%. This is a wide interval relative to the 6% case fatality observed in the sample. The case fatality could in fact be nearly double of what was actually observed. Compare it with the interval (53%, 75%) obtained earlier for cases with poor prognosis. This interval is narrow relative to the 64% rate observed in the sample. In general, the CI is narrow relative to p when p is around 0.5, between 0.2 and 0.8, and wide relative to p when p is either very low or very high.

CI for mean μ. Obtain 95% CI for μ as $[\bar{x}-2\text{SE}(\bar{x}), \bar{x}+2\text{SE}(\bar{x})]$ This gives the

$$95\%\ \text{CI for}\ \mu\ (\sigma\ \text{known}):\ (\bar{x}-2\sigma/\sqrt{n},\ \bar{x}+2\sigma/\sqrt{n}) \qquad (6)$$

However, σ is rarely known. It is then replaced by the sample SD s. Because of this replacement, the Gaussian distribution needs to be replaced with Student's t-distribution. Thus

$$95\%\ \text{CI for}\ \mu\ (\sigma\ \text{not known}):\ (\bar{x}-t_\nu s/\sqrt{n},\ \bar{x}+t_\nu s/\sqrt{n}), \qquad (7)$$

where t_ν is the value of t at ν degrees of freedom from standard probability tables. This corresponds to the probability column 0.025 so that the total probability outside $\pm t_\nu$ is 0.05. In this case, $\nu=(n-1)$. For a large n, t can be approximated by Gaussian z but there is no need to do this since t-tables are easily available.

Proportion and mean are the two most common indices on which CI are drawn. There might be isolated examples where the interest is in CI for median or for a decile, even σ. The basic method to obtain a 95% CI is to get the 2.5th and 97.5th percentiles of the distribution of the corresponding 'statistic' in the sample.

Role of sample size in obtaining a narrow CI. In Example 5, the 95% CI for the percentage of cases with serious disease was 53% and 75%. If n changes from 80 to 500 (and nothing else changes), then the 95% CI is

$$0.64\pm 2\sqrt{0.64\times 0.36/500} \text{ or } (0.60, 0.68).$$

Thus, the percentage is sharply localized. Similar features can be demonstrated for the CI for μ also, and for that matter, for any CI. Thus, a large n helps to reach a focused conclusion.

Confidence bounds. The interest sometimes is not in the lower and upper limits of confidence but only in one of them. Such one-sided limits are called bounds. Suppose in a group of patients with heart disease on medication, the 5-year survival rate is 60%. Surgery is expensive and can be advised only if it substantially increases the survival rate. This increase could be quantified as a *minimum* of, say, 20%. This is the lower bound on the increase in survival rate in this case. Such bounds can be obtained by using one-sided probabilities in the probability tables.

Confidence intervals for differences (large n)

In many situations, the interest is in the magnitude of difference in proportions (p_1-p_2) or means $(\bar{x}_1-\bar{x}_2)$ in the two groups under study. The types of differences that are of special importance in medicine are between a placebo and a drug, between drug 1 and drug 2, between men and women, etc. CI for such differences can be obtained as follows.

Estimate the SE of differences:

$$\text{SE}\ (p_1-p_2)=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \qquad (8)$$

$$\text{SE}\ (\bar{x}_1-\bar{x}_2)=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \qquad (9)$$

The subscripts are for the two samples. When the estimate of SE is available, CI for differences can be stated as follows.

CI for the difference in two proportions: $(p_1-p_2)\pm 2\text{SE}$ as in (8),

CI for difference in two means: $(\bar{x}_1-\bar{x}_2)\pm 2\text{SE}$ as in (9)

Confidence intervals for π and μ (small n)

CI for π (small n). The 95% CI for π corresponding to different values of the observed sample proportion p can be read from the graph given in Fig. 1. This is drawn for some specific values of n. The upper and lower limits are read off the vertical axis using the pair of curves corresponding to the sample size n. The sample proportion p is on the horizontal axis. If n is not exactly as shown, visual interpolation can be done to get an approximate CI.

Example 6. **Suppose the interest is in estimating the chance of uterine prolapse in women who complain of disturbance of micturition and vaginal discharge. If $n=12$ women with such complaints could be examined and 3 had uterine prolapse, the sample proportion p is $3/12=0.25$. Figure 1 does not give the CI curve for $n=12$. However, an eyeball interpolation for $n=12$ corresponding to $p=0.25$ gives (0.06, 0.58). Thus, the chance of uterine prolapse in women with these complaints can be between 6% and 58%. The obtained CI is wide. As n increases, the CI narrows but this gain decreases with increasing n.**

CI for π when the success or failure rate is zero per cent. Consider a situation in which a surgeon did 10 operations of a

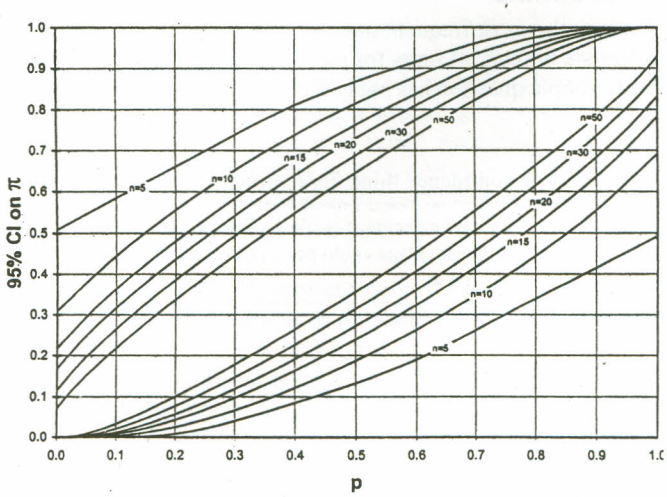

FIG 1. 95% confidence interval for π for different sample sizes

particular type for kidney stone with complete success without a single complication. Thus the complication rate is $p=0$ in this sample. Can it be concluded that the complication rate would continue to be zero for all such operations in future? Or is it just good luck in the 10 patients, who happened to be operated during that period? In statistical language, can $p=0$ be used as an estimate of π ? The obvious answer is no. The true complication rate can be estimated only by obtaining a confidence bound (CB) for π.

The 95% CB for extreme results for various sample sizes are given in Table I.

Example 7. **Consider the above example. For this surgeon, $p=0$ and $n=10$. The 95% CB for the true complication rate from Table I is 27%. Thus, we are 95% confident that the complication rate in the long run would not exceed 27%. The claim of 0% complication rate on the basis of the experience on 10 subjects is not tenable. If no complication is observed in a series of 50 such operations, then the CB corresponding to $n=50$ from Table I is only 6% ! This indicates that the size of the sample is important in determining the width of the CI, and to reach to a focused conclusion.**

CI for μ (small n—non-Gaussian distribution). We stated earlier that Gaussian form changes to Student's t-distribution when the sample estimate s is used for σ. Note that the CI in (7) is applicable only when the underlying distribution is nearly Gaussian. Many measurements done on healthy subjects follow this pattern. However, generally this is not the case with measurements in sick subjects. When the distribution is known or suspected to be far from Gaussian or of unknown shape, the methods to be used for small n are non-parametric or distribution-free methods. These methods do not depend on the exact shape of the underlying distribution.

Non-parametric methods mostly require that the values observed in the sample are ordered ascendingly $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$. Median rather than the mean is considered the central value in these methods. The CI for population median (generally denoted by $\tilde{\mu}$) is in terms of the ordered values $X_{[k]}$ and $X_{[n-k+1]}$ where k is largest, such that the probability in between these two values is at least 95%. The values of k for different n are given in Table II.

Small sample CI for differences such as $(\pi_1-\pi_2)$ and $(\mu_1-\mu_2)$ are complex. Besides these parameters, CI is also obtained in medicine and health for relative risk and odds ratio. This article does not cover these aspects.

SAMPLE SIZE

Perhaps the most frequent question faced by a statistician is the adequate number of units for inclusion in a sample. This apparently simple question has many facets. Just as a physician cannot prescribe therapy without knowing some details of the patient's ailment, a statistician cannot answer this question unless some basic information is available. The procedures to determine the sample size for estimation and testing of a hypothesis differ.

TABLE I. 95% confidence limits for extreme results

Sample size (n)	Sample percentage is 0%, true percentage could be as high as	Sample percentage is 100%, true percentage could be as low as
1	95	5
3	63	37
5	50	50
10	27	73
15	19	81
20	14	86
30	10	90
50	6	94

TABLE II. Value of k (see text) for different n—95% confidence interval (CI) for median is $(X_{[k]}, X_{[n-k+1]})$

n	k	95% confidence interval
≤5		CI cannot be computed
6	1	$(X_{[1]}, X_{[6]})$
7	1	$(X_{[1]}, X_{[7]})$
8	1	$(X_{[1]}, X_{[8]})$
9	2	$(X_{[2]}, X_{[8]})$
10	2	$(X_{[2]}, X_{[9]})$
11	2	$(X_{[2]}, X_{[10]})$
12	3	$(X_{[3]}, X_{[10]})$
13	3	$(X_{[3]}, X_{[11]})$
14	3	$(X_{[3]}, X_{[12]})$
15	4	$(X_{[4]}, X_{[12]})$
16	4	$(X_{[4]}, X_{[13]})$
17	5	$(X_{[5]}, X_{[12]})$
18	5	$(X_{[5]}, X_{[14]})$
19	5	$(X_{[5]}, X_{[15]})$
20	6	$(X_{[6]}, X_{[15]})$
21	6	$(X_{[6]}, X_{[16]})$
22	6	$(X_{[6]}, X_{[17]})$
23	7	$(X_{[7]}, X_{[17]})$
24	7	$(X_{[7]}, X_{[18]})$
25	8	$(X_{[8]}, X_{[18]})$
26	8	$(X_{[8]}, X_{[19]})$
27	8	$(X_{[8]}, X_{[20]})$
28	9	$(X_{[9]}, X_{[20]})$
29	9	$(X_{[9]}, X_{[21]})$
30+		Use Gaussian CI for mean

Sample size required for estimation

The general considerations in the estimation set-up:

1. What is the variability between subjects in the population? The sample size requirement increases as the variance increases. In the case of sample proportion p, the variance depends on the value of p. In absolute terms, a large sample is required if p or $(1-p)$ is small.
2. What is the degree of precision required? The sample size should evidently be larger when greater precision is required. In the case of proportion, the precision can be expressed in absolute terms such as a 2% difference, or in relative terms, such as 10% of p.
3. What is the least confidence in the estimate that would be tolerable? No empirical conclusion is 100% dependable but the investigator may wish to be sufficiently confident that the conclusion will be replicated in repeated samples. This requires a larger sample. If a chance of being wrong by as much as 10% or 20% is acceptable, then a small sample may be adequate.
4. Are there any subgroups of interest? If conclusions are required to be drawn for various subgroups, as in stratified sampling in some cases, then each subgroup needs to be adequately represented. This can substantially increase the total sample size.
5. How much non-response is expected? Non-response reduces the effective number of subjects that can be utilized to draw conclusions. Thus, a larger sample is required in case non-response is anticipated.
6. What sampling procedure is proposed to be used? CRS may

require a double or even larger sample size relative to SRS because of the clustering effect. Stratification and two-stage sampling may have smaller samples, if the units have been homogeneously divided.

General procedure to determine the sample size for estimation

Let the population parameter under estimation be denoted by τ and its sample estimate by t. Let δ be the difference between the two—$\delta = |\tau - t|$. Suppose the investigator requires that this difference should not exceed a specified limit L in at least 100 $(1-\alpha)$% of repeated samples. The quantity L is called precision which also is the half-width of the CI. We assume that the difference can be negative or positive. The quantity $(1-\alpha)$ is the confidence level. If confidence level $(1-\alpha)=0.95$ then the chance that the difference between the sample estimate and the parameter value exceeding the specified precision is at the most 5%. If a Gaussian form of distribution can be assumed valid, which in fact could be so in most cases when the sample size is large, then it can be shown that

$$L = z_{1-\alpha/2}\,\mathrm{SE}(t) \qquad (10)$$

where the coefficient $z_{1-\alpha/2}$ is taken from the standard Gaussian distribution. The table of probability of the Gaussian curve needs to be consulted to find a cut-off $z_{1-\alpha/2}$ such that the probability between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$ is $(1-\alpha)$. The exact value of $z_{1-\alpha/2}$ for $\alpha=0.05$ is 1.96 but is approximated to 2. For $\alpha=0.10$, $z_{1-\alpha/2}=1.645$ and for $\alpha=0.01$, $z_{1-\alpha/2}=2.58$. Thus, for a confidence level of 90% $L=1.645\mathrm{SE}(t)$, for 95% $L=2\mathrm{SE}(t)$ and for 99% $L=2.58\mathrm{SE}(t)$.

The equation (10) is the basis for the calculation of sample size for an estimation set-up. SE(t) would invariably have n in the denominator, which could then be worked out when other values are known. However, SE(t) would also contain an unknown parameter such as σ. This is to be substituted by its estimate. This could be obtained either from a previous or pilot study.

Since SE(t) always contains n, and L is specified, it is possible to obtain n from (10). This may have to be inflated to adjust for expected non-response. If estimates for various subgroups are required then this calculation has to be done separately for each subgroup. If there are many parameters under estimation, then two approaches are available. The first is to calculate the sample size for the most important parameter if that can be identified. The second is to calculate the size for all the parameters and use the one which is the maximum. The latter would give better-than-required precision on some parameters but each estimate will have at least the specified precision.

The actual formulae of sample size calculations for some estimation situations are given in Table III. These are based on the procedure explained above. The formulae are valid only for a large n, where Gaussian approximation is applicable.

Example 8. **To plan a study on the difference in the prevalence of worm infestation in agricultural and non-agricultural workers, a pilot study was carried out on 30 workers of each type. The prevalences were 33% and 20%, respectively. What sample size is needed if the difference is to be estimated within 3 percentage points with 90% confidence?**

We anticipate that the population proportions in the population would be the same as in the pilot study. Thus, $p_1=0.33$ and $p_2=0.20$. Confidence $100(1-\alpha)=90\%$ or $\alpha=0.10$. Since $d=0.03$, from the fourth formula in Table III,

$$n = \frac{z^2_{0.95}(0.33\times0.67+0.20\times0.80)}{(0.03)^2} = \frac{(1.645)^2\times(0.3811)}{(0.03)^2} = 1146$$

Thus a sample of nearly 1150 is required in each group.

Sample size for testing of hypothesis set-up with a specified power

The general considerations in the testing of hypothesis are given below. These are in addition to 1, 4, 5 and 6 stated earlier for estimation.

1. What is the level of significance required? In testing a hypothesis, the complement of confidence level is the level of significance α. A large sample is required if α is to be kept small.

TABLE III. Sample size (n) calculations for different estimations (valid for large n only; each $p \geq 0.01$)

Problem	Formula for computing n	Description of the notations
Estimating a population proportion with specified absolute precision	$\frac{z^2_{1-\alpha/2}\,p\,(1-p)}{d^2}$	*p=anticipated value of the proportion in the population d=absolute precision required on either side of the proportion
Estimating a population proportion with specified relative precision	$\frac{z^2_{1-\alpha/2}\,p\,(1-p)}{e^2p}$	*p=anticipated value of the proportion in the population e=relative precision in terms of fraction
Estimating a population mean with specified precision	$\frac{z^2_{1-\alpha/2}\,s^2}{L^2}$	s=population SD (can be estimated from a pilot study) L=specified precision of the estimate on either side of mean
Estimating a difference between two population proportions with specified absolute precision—equal n in the two groups	$\frac{z^2_{1-\alpha/2}[p_1(1-p_1)+p_2(1-p_2)]}{d^2}$	*p_1, p_2=anticipated proportions in the two populations d=absolute precision required for the difference in terms of fraction
Estimating a difference between means of two populations with specified precision—equal n in the two groups	$\frac{z^2_{1-\alpha/2}(s^2_1+s^2_2)}{L^2}$	s_1, s_2=population SD of the two populations (can be estimated from a pilot study) L=specified precision of the estimated difference on either side of the mean difference

Large n is needed so that the distribution of p can be approximated by Gaussian form * If there is no idea about the population proportion, the safest choice is $p=0.50$ For $\alpha=0.20$, $z_{1-\alpha/2}=1.28$; for $\alpha=0.10$, $z_{1-\alpha/2}=1.645$

TABLE IV. Sample size calculation for testing of different hypothesis situations (valid for large n only; each $p \geq 0.01$)—two-sided H_1

Problem	Formula for computing n	Description of the notations
Hypothesis testing for a population proportion	$\frac{[z_{1-\alpha/2}\sqrt{p_0(1-p_0)}+z_{1-\beta}\sqrt{p_\alpha(1-p_\alpha)}]^2}{(p_0-p_\alpha)^2}$	p_0=value of p under H_0 p_α=medically important value of population proportion under H_1 that is proposed to be detected
Hypothesis testing for population mean	$\frac{s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\mu_0-\mu)^2}$	s=population SD (can be estimated from a pilot study) μ_0=value of population mean under H_0 μ=medically important value of population mean under H_1 that is proposed to be detected
Hypothesis testing for difference between two population proportions—equal n in the two groups	$\frac{[z_{1-\alpha/2}\sqrt{2p(1-p)}+z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}]^2}{(p_1-p_2)^2}$	p_1, p_2=anticipated proportions in the two populations H_0:$p_1=p_2$ $p=(p_1+p_2)/2$
Hypothesis testing for difference between two population means—equal n in the two groups	$\frac{(s_1^2+s_2^2)(z_{1-\alpha/2}+z_{1-\beta})^2}{d^2}$	*s_1, s_2=population SD of the two populations (can be estimated from a pilot study) d=medically important difference between means under H_1 that is proposed to be detected

z_α is such that $P(Z \leq z_\alpha)=\alpha$. For $\alpha=0.05$, $z_{1-\alpha/2}=1.96$ and $z_{1-\alpha}=1.645$ For $\beta=0.10$, $z_{1-\beta}=1.28$; for $\beta=0.05$, $z_{1-\beta}=1.645$ If the alternative hypothesis is one-sided, replace $z_{1-\alpha/2}$ by $z_{1-\alpha}$

If a large α, say 0.10 can be tolerated, then a relatively small sample would be enough.

2. Is it a one-tail or a two-tail test? In most situations a one-tail test with $\alpha=0.05$ is equivalent to a two-tail test with $\alpha=0.10$. This is true for p-values also. Thus, one-tail testing requires a smaller sample size than a two-tail set-up.
3. How much difference between the actual value of the parameter and its value under the null hypothesis is medically important? If the minimum difference to be detected is large then a smaller sample would be adequate.
4. What is the power required? If the power is to be 99% then a bigger sample is required, compared to a power of 80%. If there is a large n, these considerations lead to an equation of the following type in most situations.

$$\text{Power}=P(Z>z_\alpha/H_1) \qquad (11)$$

The sample size n would occur on the right side of equation (11). This can be solved to obtain n when everything else is specified. Since the solution of (11) requires the actual test procedure (to be described in a subsequent article), we provide the formula of n for some specific situations in Table IV. Again, these are valid only for a large n where Gaussian approximation is applicable.

Example 9. **Suppose it is believed that the prevalence of worm infestation in agricultural workers in an area is 30%. What sample size should be chosen so that this null hypothesis is rejected with probability 0.90 if the actual prevalence is 25% or lower? Keep the probability of Type I error not more than 5%.**

In this example, $\alpha=0.05$, $1-\beta=0.90$, $p_0=0.30$ and $p_\alpha=0.25$. It is a one-sided test since the concern is with the lower values only. From the first formula of Table IV, we get

$$n=\frac{1}{(0.30-0.25)^2}(1.645\sqrt{0.30\times0.70}+1.28\sqrt{0.25\times0.75})^2=685$$

The survey should include 685 workers. If non-response is expected, the size would need to be increased accordingly.

Example 10. **Suppose a decline of at least 10 mg/dl in triglyceride levels (TGL) is considered clinically important. A group of non-vegetarian obese and non-obese subjects are proposed to be put on a vegetarian diet to see if their TGL levels decline by the clinically important magnitude. The SD of their TGL was estimated to be 15.7 and 12.5 mg/dl, respectively. This 10 mg/dl difference is desired to be detected with an 80% probability. What is the sample size if the level of significance is 10%?**

Now, $d=10$ mg/dl, $s_1=15.7$, $s_2=12.5$, $1-\beta=0.80$ and $\alpha=0.10$. Then $z_{1-\alpha}=1.28$ and $z_{1-\beta}=0.84$. With these values, the fourth formula in Table IV gives us

$$n=[(15.7)^2+(12.5)]^2(1.28+0.84)^2/(10)^2=18$$

A sample size of 18 obese and 18 non-obese subjects is enough for this study. The sample is relatively small because the difference to be detected (10 mg/dl) is quite large. If this was 5 mg/dl, then

$$n=[(15.7)^2+(12.5)^2](1.28+0.84)^2/(5)^2=73$$

Conversely, if resources permit $n=100$, then, for $d=5$,

$$100=[(15.7)^2+(12.5)^2](1.28+z_{1-\beta})^2/(5)^2 \quad \text{or} \quad z_{1-\beta}=1.21$$

From probability tables, $z_{1-\beta}=1.21$ gives $(1-\beta)=0.887$. Thus, a size of 100 will have a power of nearly 89% to detect a difference of 5 mg/dl.

REFERENCES

1 Cochran WG. *Sampling techniques.* New York:John Wiley, 1977.
2 Bennett S, Woods T, Liyanage WM, Smith DL. A simplified general method for cluster-sample surveys of health in developing countries. *World Health Stat Q* 1991; **44**:98–106.
3 Som RK. *Practical sampling techniques.* New York:Marcel Dekker, 1996.
4 Le CT. *Applied categorical data analysis.* New York:John Wiley, 1988:1–204.