
Sampling and statistical inference

Jerry W. McLarty, PhD,* and Sami L. Bahna, MD, DrPH†

INTRODUCTION

One of the major uses of statistics is drawing inferences about a population from a sample of observations obtained from the population. For example, in a clinical drug trial testing a new drug, a group of patients with a disease are randomly assigned to a new drug or an established drug. The general objective of the trial is to know how the drug will work on most or all patients with that disease, not just the particular group of patients chosen for the study. In this case, the enrolled patients comprise the *sample* from a much larger group, which is the *population* of patients from which we are sampling. If the sample is properly selected and the experiment conducted in an appropriate manner, then inferences from the sample can be applied to the population at large using a body of appropriate statistical methods. The general framework for statistical methods of inference involves hypothesis testing, that is, making a hypothesis about the experimental situation (eg, the new drug is better than the old drug) and rigorously testing the likelihood of this hypothesis using probability analysis.

SAMPLING

There are multiple methods of sampling, with each having its indications, advantages, and limitations. The primary goal is to ensure enrolling a sample that is representative of its population and free from known *bias*. Random sampling is a necessary assumption to make the statistical theory underlying hypothesis testing possible: if a random sample is assumed, we can generate statements about the data and estimate how likely our statements are to be true. Sampling by other than random means can more easily introduce bias in patient selection. Such *selection bias* could interfere with the validity of the study and limit generalization from a sample to the population. Examples of selection bias include group imbalances in sex, age, racial composition, or severity of disease. Without randomization, subtle, often unintentional, biases can occur, for example, assigning patients with the best likelihood of success to the new drug group or approaching patients most likely to be recruited for a study or most likely to comply with the protocol.

There are different methods of random sampling depending on the particular study design. Common to all random sampling methods is that chance is involved in subject selection. There are 5 common random sampling techniques: (1) simple random sampling, (2) systematic sampling, (3) stratified sampling, (4) cluster sampling, and (5) multistage sampling.

With *simple random sampling*, every possible subset or sample of subjects is equally likely to be selected. This is difficult, if not impossible, to achieve in a clinical setting because the whole population of patients with a particular disease is unknown and certainly not available for selection at the same time. For even a small population of patients, the number of possible sets can be virtually uncountable. However, what is typically done is to randomly select patients for the study from a list or take a convenience sample (eg, those who have clinic appointments during the study period) and randomly assign members of this convenience sample to the treatment arms of the study. Rather than using equally probable sets of patients, random assignment based on individual patients ensures that each patient has a known probability of being chosen. This is called *probability sampling*. Note that it is not necessary that each patient has the same probability of being in the sample just that the probability is known. For example, there may be scientific reasons to include more minorities in a study, in which case the probability of being sampled could be higher for minorities than for nonminorities.

Systematic sampling consists of sampling a proportion of a population in a systematic way, for example, randomly choosing an interval, such as every 10th patient, and randomly choosing the first patient to start with. This is an efficient means of sampling from a large population but has the danger that patient order could be manipulated by someone who knows the sampling interval. Also, it may require that the size of the population be known to choose the appropriate sampling interval.

Cluster sampling is based on similar groups or clusters of individuals, with clusters chosen randomly and individuals within the cluster used for study. An example is a nationwide study of smoking prevention and smoking cessation programs in which 11 similar pairs of communities were chosen for intervention and the intervention methods randomly assigned within pairs. Within a cluster, all the subjects may be used for study. In the paired communities study, population statistics on the health and smoking habits of each community were used as outcome variables.

One problem with random sampling is that sometimes certain groups of patients are underrepresented or an imbalance can occur between treatment arms of a study. For

Affiliations: * Department of Medicine, Cancer Prevention and Control, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana; † Department of Pediatrics and Medicine, Allergy and Immunology Section, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication January 15, 2009; Received in revised form March 12, 2009; Accepted for publication March 14, 2009.

example, a drug may have different effects on different sexes and races. With simple randomization, especially with small samples, there may be an imbalance of females assigned to the treatment arm or no African Americans assigned to the placebo arm. *Stratified randomization* is a common method to avoid unbalanced treatment assignment. Patients are stratified by factors that may have a confounding effect on the study results, in advance of randomizing. Then randomization is performed independently within each stratum. Stratified randomization has the additional advantage that at any time during the study there is a balance among strata and treatment assignment. This may be important for studies that take a long time to complete or studies that may potentially be terminated early. Note that stratified randomization is different from cluster sampling. For cluster sampling, the clusters are chosen randomly and all individuals within a cluster are studied, and often all receive the same intervention. For stratified randomization, all patients within a stratum are randomly assigned to a treatment arm.

Multistage sampling is similar to cluster sampling, except that not all subjects within a cluster are sampled. The cluster may be divided into smaller groups and random sampling used to choose the subgroups. This could be done with multiple levels. For example, clusters could be states; within a state cities could be randomly chosen, and within cities people from various neighborhoods could be randomly chosen. The disadvantage of both cluster and multistage sampling is that they typically have larger sampling errors (the sample may differ from the population) than other techniques.

All of the sampling techniques described herein depend on some kind of randomization technique. There are multiple methods of obtaining random numbers, such as from published lists or tables, pseudorandom number generator software, and even physical devices. With small samples there is a danger from using simplistic methods, such as flipping a coin, that can result in an imbalance (eg, more persons assigned to one group than the other or assigning more of the early recruited persons to one group). In summary, the point of random sampling or randomized assignment to treatment arms is to eliminate selection bias, to ensure balance in treatment arms, and to ensure that statistical methods of estimating significance are valid.

One way to reduce the risk of imbalances with small sample sizes, particularly within cluster or stratified sampling frameworks, is to incorporate *blocked randomization*. A block size is chosen, and within this block there is an equal distribution of assignment groups that are randomly assigned. In effect, this approach introduces a new layer of cluster sampling within the existing sampling framework. For example, with 2 treatments and a block size of 4, there can be an imbalance of at most 2 patients. The disadvantage of this approach is that if the treatment is not blinded, then it may be possible to predict what treatment is next when the block is almost fully allocated.

HYPOTHESIS TESTING

The scientific phases of a study involve making a hypothesis (eg, drug A is better than drug B), performing the experiment, and based on the results, then accepting, refuting, or modifying the original hypothesis. Statistical hypothesis testing is similar, with predetermined, quantified methods of testing and decision making. Perhaps unique to statistics is the use of a *null hypothesis* as the starting point for the inference process. Instead of saying drug A is better than drug B, the null hypothesis would state that “the effect of drug A is equal to the effect of drug B” or “there is no difference between the 2 drugs.” The reason for this intuitively backward way of stating the study hypothesis is that probabilities can be calculated for testing statistical significance on the condition that the null hypothesis is true. The symbol for the null hypothesis is H_0 . An example null hypothesis might be as follows:

$$H_0: \mu_A = \mu_B \text{ or equivalently } H_0: \mu_A - \mu_B = 0$$

where the mean of some outcome variable is μ_A for population A and μ_B for population B, respectively. Note that the hypothesis is stated on the population, and population parameters are by convention denoted with Greek symbols; for example, means and standard deviations are denoted by μ and σ , respectively.

We assume that the null hypothesis is true until data from our study indicate otherwise. On the basis of our experiment, we can reject the null hypothesis or not. If we reject the null hypothesis, we can calculate the probability of being wrong.

The basic algorithm for statistical hypothesis testing is as follows: (1) state a null hypothesis, (2) collect data from a random sample or randomized experiment, (3) compute a test statistic (a number that we can use to see if the null hypothesis is true or not), (4) compute a P value (a probability statement) for the test statistic, and (5) based on the P value, reject or fail to reject the null hypothesis.

The value of P used to reject a null hypothesis is set in advance of the experiment; typically $P < .05$ is traditionally considered statistically significant. The P value is the probability of rejecting the null hypothesis, by chance, when in fact it is really true (ie, the probability of a false-positive conclusion).

To illustrate the process, assume that we want to test a new drug for hypertension. In our experiment (but only after getting institutional review board approval!), we randomly assign 50 patients with hypertension to get the new experimental drug, say drug A, and another 50 patients (with comparable hypertension) to the standard of care, say drug B. For outcome measurement, we want to measure change in diastolic blood pressure after 2 weeks using the study drugs. The null hypothesis is H_0 : the mean change in diastolic blood pressure is the same for both drugs.

For a test statistic we use the familiar t test, which is used to compare 2 mean values:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{SE(\bar{X}_A)^2 + SE(\bar{X}_B)^2}}$$

where $\bar{X}_A$ and $\bar{X}_B$ are the mean changes in blood pressure for group A and B, respectively; $SE(\bar{X}_A)$ and $SE(\bar{X}_B)$ are their respective standard errors. Assume in our example we found the mean values were 20 mm Hg and 10 mm Hg for groups A and B and their standard errors were both 10 mm Hg. Then for our example,

$$t = \frac{40 - 10}{\sqrt{200}} = 2.12$$

Using probability tables or a computer program, we can determine whether the null hypothesis were true. At degrees of freedom of 98 ($n_1 + n_2 - 2$), the probability of getting a test statistic value of 2.12 or higher by chance is .02. This number, .02, is the P value.

Since .02 is less than .05 (the commonly used value for significance), we reject the null hypothesis and say the drugs are significantly different with a P value of .02. At this point, we have completed all 5 steps of the hypothesis testing algorithm.

It is possible that we are wrong and that the drugs are not really different; perhaps by bad luck we choose an atypical sample of patients. However, the statistical theory assures us that the probability of this happening is small, .02 (or 2 times of a hundred similar samples of patients).

Because the 2 groups of patients usually do not have exactly similar distributions of blood pressures, perhaps a better way of data analysis of this trial might be to compare the mean change in effect of the 2 drugs. For each drug, compute the mean difference in the diastolic blood pressure in each individual patient (ie, after – before). The comparison would be between $\bar{X}_{\Delta A}$ and $\bar{X}_{\Delta B}$.

How small a probability of error do we accept? Why is $P < .05$ used and not some other criterion such as $<.01$ or $<.10$? That is the next topic of discussion.

TYPE I AND II ERRORS

The traditional criterion of $P < .05$ to determine statistical significance is determined in advance of the experiment. It would not be fair to do an experiment and then change our criteria for significance later. Because absolute certainty is never attainable in a hypothesis testing situation, investigators must have a preset amount of uncertainty they are willing to accept. This critical value we use for significance (by convention, designated by the Greek letter α) is the predetermined probability of falsely rejecting the null hypothesis that we are comfortable with; $\alpha = .05$ is a convention, not a law. By tradition we are comfortable in taking a small risk (5%) of being in error, but there are some circumstances that would be too much risk; in other circumstances we might be willing to take an even greater risk.

Setting the critical value α is a judgment call, balancing the risk of false-positive decisions (for example, rejecting the null hypothesis when in fact there is no difference between drugs) and false-negative decisions (failing to reject the null hypothesis when in fact the drugs are different). These are

called type I and type II errors, respectively. The probability of a type I error is α ; the probability of a type II error is β . Predetermining critical values is analogous to setting the sensitivity on a smoke detector: if the sensitivity (α) is set too high, false alarms (type I errors) are likely too low and a real fire may be missed (false negative).

STATISTICAL POWER

An important consideration of designing an experiment (before hypothesis testing is done) is the probability of correctly rejecting the null hypothesis, for example, declaring a new treatment is better than the standard treatment. This probability is called *statistical power*. If β is the probability of failing to reject a true null hypothesis, then power is $1 - \beta$. Power is usually expressed as a percentage rather than a probability: 100% power is desirable but, of course, not attainable; 50% power is the equivalent of flipping a coin to make a decision. Common values for study design purposes include 80%, 90%, or 95%.

The estimation of power can be a complex statistical procedure. In general, power can be increased by the following:

- Increasing sample size; more information leads to more power.
- Increasing α ; a larger type I error increases the chance of rejecting the null hypothesis.
- Increasing precision of the outcome measure; less noise yields more power.
- Increasing the effect size; large effects are easier to find.

SAMPLE SIZE DETERMINATION

Sample size is usually the variable that can most easily be manipulated to increase the power of a study. The following simplified formula illustrates the dependence of sample size on the magnitude of the effect and the SD of the outcome measure for a 2-group comparison:

$$N = 16 \times S^2/D^2$$

where S is the standard deviation of the outcome variable, D is the difference of clinical interest in the 2 outcomes, and N is the sample size of each group. It can be seen that small differences in D would require a large sample size and that the sample size is directly proportional to the square of the standard deviation, S^2 (sample variance).

This formula is greatly simplified to be just a rough estimate of sample size but is a handy rule of thumb for many simple studies. The number 16 includes functions of the type I and II error probabilities and the normal distribution. For this formula, $\alpha = .05$ and power is 80%.

For example, the standard deviation of diastolic blood pressure is approximately 10 mm Hg. To see if a new drug is better than placebo (say, a 5-mm Hg difference in diastolic blood pressure) it would take the following:

$$N = 16 \times S^2/D^2 = 16 \times 10^2/5^2 = 16 \times 100/25 = 64$$

Sixty-four patients *in each group* should yield sufficient statistical power to detect a difference this small. However, if

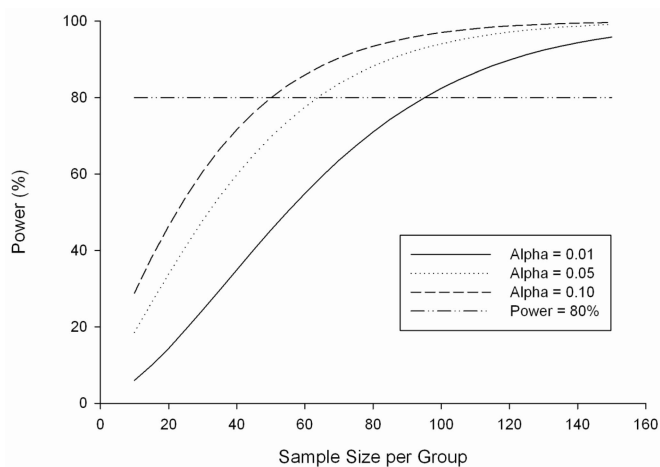


Figure 1. The effect of type I error (α) on the required sample size.

we decided that 10 mm Hg or greater would be of more clinical significance, then only 16 patients would be needed in each group. In other words, smaller differences require larger sample sizes, and larger differences can be detected with smaller sample sizes. Another caveat is that sometimes what is statistically significant may not be clinically impres-

sive. This is especially a problem with very large sample sizes, where very small numerical differences may be declared statistically significant but not be of clinical interest. For example, a drug that reduced diastolic blood pressure by 1 or 2 mm Hg might be found statistically significant with a large sample size but not be clinically relevant. Studies should be planned with the smallest practical sample size estimated to achieve clinical significance.

Figure 1 illustrates how power varies with sample size for 3 different levels of α . It can be seen that changing α from .1 to .01 almost doubles the required sample size for power of 80%.

REFERENCES

Included in a comprehensive list at the end of the Supplement.

Requests for reprints should be addressed to:

Jerry W. McLarty, PhD

Department of Medicine

Feist-Weiller Cancer Center

LSU Health Sciences Center

1501 Kings Highway

Shreveport, LA 71103-4228

E-mail: JMCLAR@LSUHSC.EDU