The Clinician as Investigator
Participating in Clinical Trials in the Practice Setting
Appendix 2: Statistical Concepts in Study Design and Analysis

Ellis W. Lader, MD; Christopher P. Cannon, MD; E. Magnus Ohman, MD; L. Kristin Newby, MD; Daniel P. Sulmasy, OFM, MD, PhD; Robyn J. Barst, MD; Joan M. Fair, RN, PhD; Marcus Flather, MD; Jane E. Freedman, MD; Robert L. Frye, MD; Mary M. Hand, RN; Robert L. Jesse, MD; Frans Van de Werf, MD, PhD; Fernando Costa, MD

I
n-depth knowledge of statistical methods is not necessary unless you are designing a trial or analyzing the data yourself, but because statistical methods are discussed in most study protocols, it is important to understand the basic terminology and concepts involved. In a randomized clinical trial, patients are randomly assigned their study treatment to avoid treatment bias, as discussed in Appendix 1 (available online only at http://www.circulationaha.org [Circulation, 2004;109:e302–e304]). Once patients are assigned to a treatment group, every effort is made to adhere to their study protocol assignments. If patients “cross over” (ie, switch treatment groups), the ability of the study to detect treatment differences will be compromised. However, for multiple reasons, switching of treatments or cessation of study treatment may occur. Because of this, most trials will analyze data with the intention-to-treat principle: To minimize any bias introduced by patients crossing over into unassigned treatment groups, each subject is analyzed as part of the group to which he or she was assigned, rather than by what therapy he or she actually received. Secondary analyses might also compare results for actual versus assigned treatment.

In clinical trials, outcomes (end points or events, such as death, need for bypass surgery, or degree of glycemic control) are considered dependent variables, whereas the intervention (the treatment assignment) is the primary independent variable. Other independent variables may include participant characteristics, such as age, smoking status, or concomitant drug therapy.

The simple reporting of sample characteristics (describing the study population, for example) or study outcomes involves the use of descriptive statistics. These generally include some measure of a central value for a collection of data, eg, a mean, or average value of the data, along with some indication of the range or spread of values. The spread can be either directly expressed as the actual range of values (eg, average age 64 years, range 18 to 81 years), or, more commonly, expressed as a mean with a standard deviation (64±25 years); the larger the standard deviation, the greater the spread of data. A central tendency can also be expressed as a median, the value in the middle of a list of sorted data.

Beyond reporting findings or describing a population, statistical methods are used to analyze study data in a comparative fashion. The most important comparisons are usually made between treatment arms of the trial. Statistical analyses are performed to determine whether there was no difference between treatment groups (the null hypothesis) or whether there was a difference between the groups larger than could be attributed to the play of chance alone (a “significant” difference).

Selection of the proper statistical tests to compare the results is crucial to the correct interpretation of the study. Generally, a study protocol that is submitted for your review will have been designed by experts in the field with the proper analyses and tests specified in advance, although it is still important to have a grasp of the statistical concepts involved as you review the proposal. If you are designing your own project, it is critical to have the assistance of a statistician, because application of the wrong statistical methods can lead to faulty trial design and erroneous conclusions.

An important concept is that of sample size, which is closely related to the power of the study. It stands to reason that the larger the number of observations (such as patients or events) in a trial, the more likely it is that a conclusion based on the trial findings will be “real” rather than due to chance,
with the results capable of being safely applied to larger populations. Hence, the larger the sample size (the larger the n), the more confident we will be in trusting the conclusion. If the effect of a particular intervention being studied is very small, a large n would be needed to be certain of detecting it. Sample size is chosen according to the desired degree of certainty that a treatment effect was not erroneously detected when there really was not an effect (a type I error) and the desired degree of certainty that a true treatment effect was not missed (a type II error). Larger sample sizes reduce the likelihood of either type I or II errors, but practicality and economics limit the size of a trial. Hence, a statement in a protocol about a study being of a particular size is arbitrary, but by convention, a probability value of less than 0.05 is generally used, which means that a statistically significant difference in outcome with an 80% probability.

Populations are often studied to determine whether there is any relationship between two or more characteristics of the population. For example, in examining a relationship between age and myocardial mass, one characteristic could be graphed against the other to determine whether as one increased, the other rose (or fell). The linear relationship between these two variables is expressed as the correlation coefficient r. If there were no relationship at all, r would equal zero. If there were a very tight relationship between the two, r would tend toward 1.0 if myocardial mass increased as age increased; if mass decreased as age increased (an inverse relationship), r would approach −1.0. A search for relationships among multiple variables uses multivariable analysis, whereas the use of data from multiple variables to arrive at a likelihood of an outcome involves multiple regression analysis.

When results from two groups are compared, it is important to be able to decide whether there is no difference between them (the null hypothesis) or whether they are really different. Various statistical tests are used depending on the nature of the data being compared, but ultimately the result of the comparison is expressed as a probability value. A probability value represents the probability that a test result would be as extreme or more extreme as the result obtained from the study, assuming that the null hypothesis was true. In other words, if a comparison yielded a probability value of 0.05, there would be a 5% likelihood that a result similar to that obtained in the study could be obtained if the two comparison groups were identical. This, in essence, is stating that there is actually a high likelihood that the null hypothesis is false and that the two groups are really statistically different. The rigor of the statistical testing insisted on to reject the null hypothesis is arbitrary, but by convention, a probability value of less than 0.05 is generally used, which means that a statistically significant observation has no more than a 5% chance of a type I error (rejecting the null hypothesis when it is really true, i.e., of finding a difference between two groups when there really is none). The Table shows a table typical of those used in comparing patient characteristics in a cardiology trial, with probability values in the rightmost column to provide a measure of the significance of the comparisons. Practically speaking, it is important to note that although the difference between two groups may be statistically significant to a very small probability value, the difference may be of no clinical significance. For example, a trial may determine that pulmonary arterial wedge pressures were 22 mm Hg in a treatment group and 25 mm Hg in the control group, with a probability value of less than 0.05; the difference in this case is statistically significant but not clinically relevant.

Data that reflect time to an event, such as time to hospitalization or time to death, are often reported with survival curves, also known as Kaplan-Meier curves, because it is more difficult to compare these data with means or the tests discussed above. Because not every subject may have experienced the event of interest by the end of the study, data for these individuals are considered to be censored at their last known date of follow-up. Survival analysis methods uniquely

<table>
<thead>
<tr>
<th>Sample Table of Patient Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The P values in this table show that Group A had significantly fewer women, and patients in this group were significantly older. However, differences in diabetes, hypertension, and previous infarction were not statistically significant. The P value for the “previous angioplasty” comparison was not significant (0.0768), but P values close to 0.05 are sometimes described as “borderline significant” or “approaching significance,” or the information may be described as “Patients in Group A had a trend toward less previous angioplasty.” This type of notation should be used with caution.

![Figure 1. Survival curve. Dashed and solid lines represent control and experimental groups, respectively; probability value for comparison is shown.](image-url)
address time-to-event data for which censoring may be present. Survival curves look like downward-sloped stepped staircases (Figure 1); statistical analysis is performed to determine whether there are significant differences between the two curves.

Comparative data between two groups are sometimes expressed as a ratio of the risk of an event occurring in the two groups being examined (relative risk, or odds ratio). If the risk of the event is the same in the two groups, the ratio, or relative risk, is 1.0. Data expressed as relative risk also include a range of values, usually calculated so that the actual relative risk falls within this range 95% of the time. This is called a 95% confidence interval. Figure 2 shows relative risk as it is often displayed graphically, as an odds ratio plot or “box-and-whisker” plot.

Related to the relative risk of an event between two groups is the concept of number needed to treat. If a particular therapy resulted in an absolute difference in mortality rates of 50% between the treated and untreated groups, 2 patients would need to be treated with that therapy to prevent 1 death. Clearly, the smaller the advantage conferred by a treatment, the larger the number of patients needed to treat before an improved outcome is seen.

For a more in-depth review of statistical issues, a number of excellent references provide a good basis. 1–3 A useful glossary of statistical terminology is found in each issue of the American College of Physicians’ Journal Club, as well as in each issue of Evidence-Based Medicine.

References
The Clinician as Investigator: Participating in Clinical Trials in the Practice Setting:
Appendix 2: Statistical Concepts in Study Design and Analysis
Ellis W. Lader, Christopher P. Cannon, E. Magnus Ohman, L. Kristin Newby, Daniel P. Sulmasy, Robyn J. Barst, Joan M. Fair, Marcus Flather, Jane E. Freedman, Robert L. Frye, Mary M. Hand, Robert L. Jesse, Frans Van de Werf and Fernando Costa

Circulation. 2004;109:e305-e307
doi: 10.1161/01.CIR.0000129020.43304.A7
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2004 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/109/21/e305