SUPPLEMENT

Supplement to:

ANNALS OF
Allergy, Asthma
& Immunology

October 2009 • Volume 103, Number 4, Supplement 1

Statistics for Clinicians

Editor:
Sami L. Bahna, MD, DrPH

~Official Publication of~

ACAAI
American College of Allergy, Asthma & Immunology

Statistics for Clinicians

Sami L. Bahna, MD, DrPh

TABLE OF CONTENTS

INTRODUCTION

After my medical school graduation, I aspired to pursue an academic career. The only appointment available to me was at an institute of public health. During the 5 years of preparation for my doctorate degree in public health, I enjoyed gaining new knowledge of biostatistics and epidemiology, but I missed direct patient care. I got special pleasure during the performance of my thesis study on the epidemiology of allergic disorders in school children, which incidentally was just after the discovery of IgE.

On immigration to the United States, I decided to pursue a career in clinical medicine. During my pediatrics residency and allergy/immunology fellowship, I was able to perform and publish a modest number of studies—primarily because of my applying the principles of study design and data collection, analysis, presentation, and interpretation. I came to realize the value of my initial "detour" into the field of public health, which continued to facilitate my academic career in clinical medicine.

Even though I did not keep up with the advances in statistical methods and computer-assisted programs, my modest basic knowledge in statistics has been of significant help. It is unfortunate that medical training does not allow time for such an important subject, of value not just to investigators but to readers of the scientific literature at large. Books on statistical methods are numerous but mostly voluminous and usually written by experts outside the biomedical field. I became preoccupied with a desire to prepare simplified information on the subject. My interest became enhanced through encountering some colleagues at my current institution who are giving scattered lectures on statistical methods to residents and fellows. Their knowledge and expertise far exceed mine, particularly in the use of new technologies. We organized a weekend course on statistics at our institution. It was very well attended, surprisingly, more by members of the faculty, who were required to pay fees, than by residents and fellows, whose registration fees were waived.

My desire to prepare a simplified primer on statistics became stronger while serving as a reviewer for journals and on a few editorial boards. When I became Associate Editor of the *Annals of Allergy, Asthma and Immunology*, I mentioned my desire to the Editor-in-Chief, Gailen Marshall, MD, PhD, who was very receptive and encouraging. This supplement to the *Annals* was prepared in collaboration with my colleagues Steven Conrad, MD, PhD, Jerry McLarty, PhD, and Runhua Shi, MD, PhD, who have admirably presented complex topics in a simpler and clearer way. I am very grateful for their contributions. I hope that this supplement will help medical trainees and clinicians of various specialties in their interpretation of published data and their performing of research studies.

In addition to my gratitude to my friend, Dr Gailen Marshall, I thank the *Annals*' editorial staff, particularly Ms Laura King. I would also like to express my appreciation to my colleagues in the leadership of the American College of Allergy, Asthma & Immunology (ACAAI) for making Statistics for Clinicians available to so many by its publication as a supplement to the College's official journal.

Sami L. Bahna, MD, DrPH, ACAAI President (2009–2010), Department of Pediatrics, Allergy and Immunology Section, Louisiana State University Health Sciences Center, Shreveport, Louisiana, SBAHNA@LSUHSC.EDU

Foreword

Statistics for Clinicians

"The data must be valid since the *P* value is less than .05." Some form of that statement has likely been made (or at least thought) by most of us somewhere along our academic journey as we viewed presentations or publications that made rather strong claims about medication or procedural efficacies, risks, pathophysiologic differences, and other factors. But just what does a *P* value *really* mean? And can something be statistically significant but clinically trivial? Was the sample size sufficient to draw clinically useful conclusions?

As the Editor-in-Chief of the *Annals*, I am privileged to lead an extraordinary team of associate editors and reviewers who address these questions every day. As a clinical immunology and allergy division director responsible for an accredited allergy/immunology training program, I see fellows struggle to understand how to optimally design an experiment, calculate a valid sample size, and select the statistical tools they plan to use *before* they start an experiment. As a researcher, I am constantly on guard during post hoc analysis of data to attempt to minimize bias that affects the conclusions that we draw from our data.

There are many biostatistics textbooks and courses available to guide the proper use of statistical techniques. There are even some simple "how-to" guides to "teach" statistics to the uninitiated. However, the utility of such tools is often limited unless one has access to a full-time biostatistician—someone to whom our clinician readers in practice and even fellows in many training programs often have limited access.

When I was approached by one of our associate editors, Dr Sami Bahna, about the possibility of putting together a work that would be a resource in biostatistics, I was delighted to help him meet the need by making this volume a reality for our readers. He assembled an extraordinary team of statisticians, who have compiled a scholarly, easy to understand, and, perhaps most importantly, *practical* guide to statistics. This work will be of considerable use to practicing clinicians and fellows in training as they critically read the literature that will guide their future practice patterns and investigators, both basic and clinical, as they generate new knowledge in a scientifically and statistically valid fashion that will move our subspecialty forward to better treatments for our patients.

The *Annals* is honored to host the publication of this supplement. I extend my thanks to the authors, in particular Dr Bahna as guest editor, for their skilled writing and composition of this tome. It should be of considerable value to clinical medicine as a whole for some time to come.

GAILEN D. MARSHALL, MD, PHD
Editor-in-Chief

Research study design

Sami L. Bahna, MD, DrPH,* and Steven A. Conrad, MD, PhD†

INTRODUCTION

Study design is the most important first step in a research project. Comprehensively, a research study comprises several components, starting with stating clear objectives and ending with conclusions and recommendations (Table 1). In general, studies may be descriptive or analytical. *Descriptive studies* provide description of a condition in one or multiple persons. *Analytical studies* are based on certain hypotheses, follow a specific protocol, and involve measurement, classification, and statistical analysis of data. Research studies may be classified into 2 main types, namely, epidemiologic and interventional (Table 2).

Table 1. Components of Research Studies

State clear objective(s)
Delineate a hypothesis
Define the study population
Plan a sound study design
Choose appropriate methodology
Decide on a minimal valid sample size
Data collection and organization
Data analysis and presentation
Interpretation and discussion of findings
Conclusion(s) and recommendation(s)

Table 2. Types of Research Studies

I. Epidemiologic studies (observational)
A. Descriptive
B. Cross-sectional (survey)
C. Prospective (longitudinal, cohort, or follow-up)
D. Retrospective (case-control)
II. Interventional studies (clinical or experimental)
A. Therapeutic trials
B. Secondary prevention trials
C. Primary prevention trials

EPIDEMIOLOGIC STUDIES

Descriptive Studies

In a descriptive study, the author describes the characteristics of affected persons in a sample of one, a few, or a series of subjects. Observations may include age, sex, occupation, geographic distribution, physical findings, comorbid conditions, specific associated factors, and others. This type of study is easy and quick, but the provided information is largely left up to the authors, and hence it can be incomplete or biased. Statistical analysis would depend on the data collected but typically includes descriptive and comparative statistics.

Cross-sectional Studies

A cross-sectional study is the most common type of epidemiologic study. It is usually intended to answer a question of interest through collecting data on a population sample (a cohort) and its exposure factors. Criteria are specified for the target and accessible populations, and then appropriate methods are used for drawing the sample. The data to be collected should be tailored toward answering the research questions. The measurements are made once without a follow-up. It is a snapshot of the condition as it exists in a certain sample at a particular point of time or within a short interval. The cohort sample size should depend on the frequency of the disease in that population and the precision of estimate required (ie, the lower the prevalence, the larger the sample needed; see the article entitled "Sampling and Statistical Inferences" in this issue).

Cross-sectional studies reveal prevalence, which is the number of cases at one point in time divided by the number of people at risk at that time. They have the advantages of being relatively fast and inexpensive. In addition to describing the demographic and clinical characteristics of the study group, they may reveal possible associations and generate hypotheses for further studies. However, causal relationships may be difficult to establish by cross-sectional studies. Another limitation is that they do not provide worthwhile information on the prognosis or the natural history of the disease. Serial cross-sectional studies on a particular population would be valuable in providing information about changing patterns of the disease over time.

Prospective Studies

Prospective studies are also called longitudinal, cohort, or follow-up studies. In this type of study, an attempt at answering the questions of interest is performed by following up a group of initially nonaffected subjects longitudinally to document the development of the condition under study and the possible causative factors. The subjects being followed up may be enrolled as a cohort rather than individually in a sequential manner. Follow-up can be periodic during a certain interval or at the end of the study duration. Prospective studies may or may not include intervention. In the latter case, a specific intervention (exposure or treatment) is ap-

Affiliations: *Department of Pediatrics, Allergy and Immunology Section, Louisiana State University Health Sciences Center, Shreveport, Louisiana; †Departments of Medicine, Emergency Medicine, and Pediatrics, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication February 4, 2009; Received in revised form March 27, 2009; Accepted for publication March 30, 2009.

plied to one cohort and comparison is made with another similar cohort without intervention. Prospective studies have several advantages and disadvantages (Table 3). The most important advantage is the collection of complete and standardized data. The main disadvantages are the cost and long duration of follow-up. Also, outflux or influx in the population or changes in exposure can impose problems in data analysis or introduce bias.

Retrospective Studies (Case-Control)

In retrospective studies, a group of cases of disease is identified and its exposure history is compared with that of a control group from the same population. Retrospective studies are most desirable for studying diseases that have low prevalence or that require a long induction or latent period to develop, when exposure data are difficult to obtain, or when the study population is dynamic. Case-control studies can reveal factors that lead to the occurrence of disease and estimate the risk and odds ratios. Their greatest advantages are saving time and cost. A great limitation of retrospective studies is recalling of past events or gathering data from different sources that can vary in completeness or accuracy (Table 4).

The definition of *disease* or *case* has a great impact on the study findings. Validity of the results would depend on the criteria being applied (ie, specific vs nonspecific and objective vs subjective). For example, a case of asthma in different studies may be based on any of the following: self-diagnosis, history of recurrent wheezing, diagnosis by any health care professional, diagnosis by a specialist, response to bronchodilators, or specific pulmonary function test results. Inclusion and exclusion criteria should be explicit and clear. Restrictive criteria, though enhancing the purity of data, would reduce the number of subjects. This issue is of great importance in comparing the results of different studies. Once a case definition is settled, subjects can be identified through various database sources known for their high accuracy and efficiency (eg, clinics, hospitals, special registries). In some instances, advertisements in media or in selected common places can identify persons for screening.

The control group should be sampled independently of the case group and must represent the source population. It can be difficult to enroll controls for retrospective studies because they do not have the same level of interest or the recalling of past events as the cases do (recall bias). They should be selected randomly from a well-defined population or sources that guarantee no or minimal bias. The controls should be matched with the cases regarding confounding factors that might affect the development of the disease. The size of the control group should preferably be at least equal to the case group. Increasing the number of controls has the advantage of slightly increasing the power of the study, but a ratio of up to 4 controls to 1 case would be sufficient to avoid increasing the cost or duration of the study. Sometimes more controls than cases are used to ensure generalizability of results or to reduce the likelihood that unknown hidden variables may be responsible for outcomes.

Table 3. Advantages and Disadvantages of Prospective Studies

Advantages	Disadvantages
Uniformity and completeness of data	Require large sample size
Calculation of absolute and relative risk	Long follow-up periods
Can have multiple end points	High costs
Can study multiple outcomes	Dropouts
Provide incidence	

Table 4. Advantages and Disadvantages of Retrospective Studies

Advantages	Disadvantages
Short duration	Inaccuracy or incompleteness of data because of the recall factor
Low cost	Heterogeneity in data sources
Calculates odds ratios	May not identify weak exposure factors
	Susceptibility to sampling bias of using 2 populations
	Limited to one outcome measure

INTERVENTIONAL STUDIES

Interventional studies are typically either therapeutic or preventive in which a particular procedure or measure is being evaluated regarding its efficacy and safety. Clinical or therapeutic trials are intended to test the effect of a certain treatment on the disease outcome. *Secondary prevention trials* are intended to test the effect of a certain intervention in preventing the recurrence or complication of a disease. *Primary prevention trials* are performed on apparently healthy subjects with the objective of preventing the development of the disease. To increase yield and cost-effectiveness, this type of trial is usually performed on subjects at high risk of developing the disease.

Subjects to be included in interventional trials are enrolled according to clearly stated predetermined inclusion and exclusion criteria. In addition to voluntary withdrawal by the subject at any time during the study, there should be certain criteria for withdrawing the subject before completing the trial. It is important to follow up all patients who entered a trial, even if they withdraw from the study because exclusion of these subjects may introduce bias. Each trial may have one or more well-defined end points and outcome measures, preferably including objective ones. Limiting the number of end points makes follow-up easier and also reduces the chances of bias due to variations in the follow-up period among subjects.

Crossover Designs

In studying chronic diseases, a certain therapy may be compared with another while the subject serves as his/her own control. Patients are randomly assigned to receive 2 types of therapy in different sequences (ie, first treatment A then B or first treatment B then A). The duration between the 2 treat-

ments (washout period) would depend on the duration of effect of the drug therapy after its discontinuation (ie, carryover effect). The period of treatment with each drug depends on the expected duration until maximal response. The major limitation of crossover studies is the uncertainty of whether there is a carryover effect or not, and if so how long the washout period should be.

Factorial Designs

Interventional studies may need to address the effect of more than one treatment regimen, each may include more than one therapeutic agent or procedure. For example, a trial on patients with uncontrolled asthma may be designed to see whether the addition of an oral antihistamine (anti-H_1) or of an oral leukotriene antagonist (LTA) to an inhaled corticosteroid (IC) would have a beneficial effect. The subjects should be randomly assigned to 1 of 4 groups:

A. IC only
B. IC plus anti-H_1
C. IC plus LTA
D. IC plus anti-H_1 plus LTA

To avoid bias on the subject's part, every patient in each group should receive 3 "medications." Therefore, group C should receive a placebo in lieu of anti-H_1, group B receives a placebo in lieu of LTA, and group A receives 2 placebos: one for anti-H_1 and another for LTA. In this example, the effect of anti-H_1 is addressed by comparing group B with group A, that of LTA by comparing group C with group A, and that of the combination of anti-H_1 and LTA by comparing group D with group A. These comparisons are not performed individually but as part of an overall factor analysis.

In designing clinical trials, provisions for intervention should be in place for subjects whose disease deteriorates during the trial. Subjects may be withdrawn from the trial or may crossover from a treatment arm to the control arm if the latter comprises standard of care. In an intent-to-treat analysis, subjects removed from their initial assigned group are analyzed as if they remained in the group.

Control and Placebo Groups

Comparative therapeutic trials test the efficacy of one or more treatments against a control group. The control group is the benchmark against which an improvement is sought. The preferable type of control group is the *placebo* or an inactive treatment. A finding of efficacy is then conclusive evidence that the treatment has an effect. Although placebo groups are preferred, there may be ethical reasons to use an *active control*, which is an existing, approved, or widely accepted treatment that, if withheld, would deprive the subject of standard care. Although the trial may be designed to demonstrate efficacy over that of the active control, a more common approach is a *noninferiority design*, with the goal of demonstrating efficacy equal to that of an accepted therapy. The major pitfall of noninferiority designs is that the accepted therapy used as an active control may not have achieved its status as standard of care through rigorous trial design, and therefore demonstration of noninferiority would not necessarily equate to efficacy.

Intent to Treat

In any clinical trial, some subjects who have been enrolled in the study and have been randomized to receive the treatment may drop out of the study at any time. Or, some patients in the placebo arm of the study may start taking the active drug on their own; these are called *drop-ins*. In either case, it may be preferable to compare the intent to treat of one group with the intent to treat of another group (ie, proceed with the analysis as if no patient dropped in or out). Exclusion of such patients not only reduces the sample size but may introduce bias, particularly if the withdrawal was because of a negative outcome during the treatment. However, if some patients after enrollment were found to be ineligible for the trial, they should be excluded.

QUESTIONNAIRE DEVELOPMENT

Using questionnaires is a common way of gathering uniform data in clinical studies. They may be completed by the responder (self-administered), by an interviewer, or by a combination. Both methods are susceptible to errors imposed by the memory effect and the tendency of responders to provide socially acceptable answers. The advantages and disadvantages of each method are summarized in Table 5. Telephone interviews can reduce the cost but may be difficult to conduct at an optimally convenient time to the subject or to the staff.

General Principles of Questionnaire Design

Designing an attractive, easy to complete questionnaire is of utmost importance in increasing the rate of response and the completeness and accuracy of data.

- At the beginning, include a brief description of the purpose of the study and how the data will be used.
- Provide simple clear instructions on how the questionnaire be filled.
- Begin with identification data: name (unless anonymous), age or birth date, sex, address (or geographic location), method of contact (eg, telephone, e-mail, fax), and date of completing the questionnaire.
- The number and contents of questions depend on the data that need to be collected. Questions relevant to a certain issue should be grouped together under a heading or short statement.

The questions may be constructed in 2 ways. Open-ended questions have the advantage of obtaining answers in the responder's own words but the disadvantage of difficulty to code or group so that statistical analysis is feasible. Closed-end questions ask the respondent to choose from preselected, mutually exclusive potential answers, which saves time and allows easy grouping and analysis. This method has the disadvantage of limiting the answers to the investigator's choices and may not reveal important unexpected responses. This can be overcome by adding a category of "other (specify)." Providing only 2 potential answers may be suitable for

Table 5. Comparison Between Self-Administered and Interviewer-Administered Questionnaires

Type	Advantages	Disadvantages
Self-administered	Low cost Can be anonymous to protect privacy and encourage honest answers Exclude potential bias by interviewers Mailed questionnaires allow convenience for responders and time to recall or gather information	Potential misunderstanding of questions Incomplete or inaccurate data Poor handwriting Low return rate Influenced by the subject's degree of literacy
Interviewer-administered	Interviewer can clarify questions Low effect of subjects' literacy level High degree of accuracy and completeness Allows interviewer to collect observational information or administer certain tests	Does not protect privacy Responders may provide socially acceptable answers resulting in inaccurate data Very personal questions may not be answered Require more staff, time, and cost

certain qualitative variables but may not be optimal for quantitative variables. The responder may be asked to choose only 1 answer or all that apply or to give a rank. Certain questions may include the option of "nonapplicable" or "do not know." Questionnaires should be designed such that all questions are answered or denoted as not applicable or not answered. It is essential to data analysis to know whether a data value is missing or not. In some situations, blanks may be erroneously interpreted by statistical software as having a value of 0.

- Questions should be constructed with utmost neutrality, clarity, and simplicity.
- Borrow from others (ie, take advantage of questionnaires used in previous similar studies with the privilege of modifications).
- Quantitative categories should encompass the whole anticipated range, meaningfully classified, and should be exclusive (ie, the beginning of a category should not be the end of the preceding category).
- The sequence of potential responses should be in a logical order to facilitate answering and later coding.
- The questionnaire should be reviewed by experienced colleagues, including the statistician who will be involved in the data coding and analysis.
- Pretesting should be performed first on a small number of potential responders and after revision on a larger number. This process may result in clarification, estimation, or addition of questions.
- Interviewers should be selected for high skill and should receive training on administering the questionnaire to maximize the accuracy and minimize introducing bias.
- The final findings' reliability largely depends on the accuracy and completeness of data collected. Methods for handling missing values should be explicit in advance of data collection.

CLINICAL TRIAL PHASES

Development of new drugs begins in the laboratory through biochemical procedures and animal experiments. Once those preclinical tests show promise of efficacy and probable safety, human trials can be performed in phases, usually 3, before their approval for marketing. A postmarketing fourth phase may be performed for additional purposes. Study design parameters, such as the need for control groups, randomization, and sample sizes, may differ considerably for each of the 3 study phases.

Phase 1

This preliminary trial is performed on a small number of patients (usually 20 to 30) primarily to determine the drug's safety. It involves frequent visits for recording symptoms, physical findings, and laboratory tests that may include complete blood cell count, liver function, renal function, electrocardiogram, etc. Different dosage regimens may be tested to determine the dose that causes minimal adverse effects. This phase is most commonly performed on healthy male volunteers and in addition to safety testing includes pharmacokinetic and pharmacodynamic studies in humans. Additional phase 1 trials may be performed on the target population (those with the disease to be treated) because the drug's pharmacokinetics and pharmacodynamics and safety may be altered in this group. In the case of toxic treatments, for example, cancer chemotherapy or radiation, it is unethical to use healthy volunteers, so actual patients with the disease must be the study subjects.

Phase 2

Once phase 1 shows an acceptable safety profile, the drug can be tested in a larger number of patients (usually >100) primarily for efficacy but also for safety. In this phase, the drug may be compared with a placebo or a current standard treatment in a double-blind fashion (ie, neither the patient nor the assessing investigator is aware of the type of treatment). This phase also involves frequent visits and laboratory tests but often less than what was recorded in the phase 1 trial. The sample size is typically small, and phase 2 findings are used to support the decision to move to phase 3 trials for purposes

of approval. There may be more than one treatment arm to further evaluate optimal dosing if phase 1 data are not conclusive.

Phase 3

After documenting a statistically significant efficacy of the new drug (usually above placebo), the drug is tried in a larger number (hundreds or thousands) of patients, primarily focused on efficacy while continuing to monitor safety. A head-to-head comparison is performed against a placebo control or a currently accepted treatment (active control) if a placebo would be unethical, usually in equal proportions. Monitoring this phase may include fewer parameters than in phase 2, unless findings from phase 2 warrant closer monitoring. Patients are randomly assigned to either treatment or control in a double-blind fashion. One, or typically more, of these phase 3 (registration) trials is performed for drug approval. In the United States, the decision for approval is made by the Food and Drug Administration based on a sound study design, significant efficacy, and an acceptable safety profile.

Phase 4

After marketing of the drug for some years, the manufacturer may choose to perform postmarketing clinical trials. The objective of the trial may be to compare the drug with other available treatments, to investigate the efficacy and safety of new dosage or combination regimens, or to explore new potential indications for the drug.

BIAS AND ITS SOURCES

Bias has a systematic influence on the observations in a trial. Bias can occur at any stage of the study and can have various sources. Its direction may be toward or away from the null hypothesis. Its impact would depend on its degree. Every effort should be made to prevent it through a sound study design and careful performance of the study, beginning with the selection of the subjects, collection of measurements, and data analysis, presentation, and interpretation. Bias is of 2 main types: selection and observational.

Selection bias can arise from differences in selecting or following up the study groups. Therefore, it can occur from inappropriate selection of cases or controls, self-selection of subjects, or differential loss to follow-up.

Observational bias arises from differences in the way of collection of data. It can lead to enrolling subjects whose disease or exposure is incorrect. Its sources can be poor recall of past information, suboptimal interview method, or interviewer's error in obtaining measurements or data collection. Observational bias can be prevented by ensuring impartiality of the interviewers and the study subjects, appropriately designing the questionnaire or data collection form, and ensuring high accuracy of data.

REFERENCES

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

Requests for reprints should be addressed to:
Sami L. Bahna, MD, DrPH
Department of Pediatrics
Allergy and Immunology Section
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-3932
E-mail: SBAHNA@LSUHSC.EDU

Descriptive statistics

Runhua Shi, MD, PhD, and Jerry W. McLarty, PhD

INTRODUCTION

Statistics can be thought of as 2 distinct activities: descriptive statistics and inferential statistics. As its name implies, descriptive statistics describe the characteristics of data. Inferential statistics on the other hand take a more experimental approach to the analysis of data and consist of testing hypotheses about the data and inferring properties of a population from samples. In this article, the basic concepts of statistics, type of distributions, and descriptive statistics will be presented. A few examples will be provided to illustrate the concepts.

BASIC CONCEPT OF STATISTICS

Statistics

Statistics is the field of study concerned with the collection, organization, summarization, and analysis of data and the drawing of inferences about a body of data when only a part of the data is observed. Biostatistics is application of statistical tools in the field of biological sciences and health.

Descriptive Statistics

Descriptive statistics is the information used to describe the data or statistics, such as the average values of the data and how variable they are and what shape the distribution of data takes.

The raw material of statistics is called data. The data are the numbers that result from measurements or counting. The sources of data are included but not limited to the following: routinely kept records, survey results, experiment results, and electronic databases.

It is necessary to introduce some basic concepts related to statistics before the statistics for different distributions are discussed. These concepts include the definitions of a variable, a random variable, measurement scale, quantitative variables, and qualitative variables.

A *variable* is an observable characteristic that takes on different values for different people, places, or objects. For example, age is a variable and it can range from 0 to 120 years.

A *quantitative variable* is a variable that is measured and conveys information regarding amount. For example, age is a quantitative variable that can have a numerical value.

A *qualitative variable* is a variable that conveys information regarding attribute. These attributes can be counted. For example, sex is a qualitative variable; it has a value of male or female.

A *random variable* is a variable whose values arise as a result of chance factors and cannot be exactly predicted in advance. For example, age can be a random variable in a study if each person's age is not known in advance. If a random variable is a qualitative variable, then we call this a *discrete random variable* (ie, a random variable that is characterized by gaps or interruptions in the values that it can assume; eg, male or female).

If a random variable is a quantitative variable that does not possess interruptions or gaps, we call this a *continuous random variable*. It can take any numerical value, including a fraction.

A *population* of entities is defined as the largest collection of entities for which we have an interest at a particular time, for example, the whole US population (300 million, all patients with asthma, and all children younger than 10 years). A *sample* is a smaller subset of individuals that we may have access to for study. For example, people in one state, patients with asthma who are seen at one medical center or hospital, and children younger than 10 years who are patients in a particular clinic.

MEASUREMENT AND MEASUREMENT SCALE

Measurement

Measurement is the assignment of numbers to objects or events according to a set of rules. This is important statistically because the type of data can determine the type of descriptive statistics or analytical techniques to be used.

The Nominal Scale

The nominal scale consists of naming observations or classifying them into various mutually exclusive and collectively exhaustive categories. For example, sex can be measured on a nominal scale as male and female.

The Ordinal Scale

The ordinal scale consists of ranking observations according to some criterion. For example, education level can be an ordinal scale from grade 1 to grade 12. However, with some ordinal data there is no assumption about distance between levels. For example, variables with values such as mild, moderate, or severe might be coded as 1, 2, and 3, respectively. However, it cannot be assumed that moderate (2) is twice as bad as mild (1) or that the difference between 1 and 2 is the same as the difference between 2 and 3.

The Interval Scale

The interval scale consists of ordered measurements with a well-defined distance between 2 measurements; for example,

Affiliations: Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication June 7, 2009; Received in revised form June 8, 2009; Accepted for publication June 8, 2009.

body temperature of 98.6°F^0 or 37°C^0. For both temperature scales, the definition of 0 is arbitrary.

The Ratio Scale

The ratio scale consists of using a measurement scale that contains a true zero point and can be expressed as the ratio of 2 numbers. Ratio data can have an infinite number of values. Weight, blood pressure, and serum cholesterol level are examples of ratio level data.

DESCRIPTIVE STATISTICS

Descriptive statistics are used to describe the basic features of the data gathered from an experimental study in various ways. They provide simple summaries about the sample and the measures. The measures of location, spread, or dispersion are frequently used statistics.

The *measures of location* describe where the center, middle, or most of the data are located. The arithmetic mean, median, mode, and geometric mean are commonly used location measures.

The *arithmetic mean* is one of the measures for location and is the sum of all observations divided by the number of observations. The mean value is what is referred as the *average*.

If we denote a set of data by $X = (x_1, x_2, \ldots, x_n)$, then the sample mean is typically denoted with a horizontal bar over the variable:

$$\overline{X} = \frac{1}{\mathbf{n}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{X_i}.$$

For example, a data set includes these values: 1, 10, 100, 1,000, 10,000, and 100,000. The arithmetic mean is (1 + 10 + 100 + 1,000 + 10,000 + 100,000)/6 = 18518.5.

The *median* is the middle value of the series of measurement. The sample median can be derived as when n is odd: the median is the $[(n + 1)/2]^{th}$ observation. For example, there are sorted 9 records about worker's age; these are 23, 23, 24, 28, 30, 40, 43, 44, and 48. Because n = 9, the median is (9 + 1)/2 = (10/2) = 5th observation (median = 30). The number of observations below the age of 30 years is the same as the number of observations above the age of 30 years.

When n is even, the median is the average of the (n/2)th and $(n/2 + 1)^{th}$ observation. For example, there are sorted 10 records about worker's age; these are 23, 23, 24, 28, 30, 38, 40, 43, 44, and 48. Because n = 10, the median is the average of 10/2 = 5th and 10/2 + 1 = 6th observations; it is (30 + 38)/2 = 34. The median differs from the mean in being not affected by extreme values. For example, the median of 3 numbers, say, 3, 5, and 7, is 5, but the median of 3, 5, and 70 is also 5. This is often useful when data are nonsymmetric or skewed.

The *mode* is the most frequently occurring value among all the observations in a sample. For example, there are 10 records of age in a study, such as 23, 23, 24, 28, 28, 28, 30, 40, 43, and 44. The mode is 28. A certain variable may have more than one mode.

Geometric mean may be appropriate to show the location of the data that are not normally distributed. The geometric mean ($\overline{X}_g$) is computed by the following formula:

$$\overline{\mathbf{X}}_\mathbf{g} = \sqrt[\mathbf{n}]{\mathbf{X}_1,\ldots,\mathbf{X}_\mathbf{n}} = \prod_{\mathbf{i}=1}^{n} \mathrm{X}_n^{(1/\mathbf{n})}$$

Logarithmic identities can be used to transform the formula as following:

$$\log_{10} \overline{\mathbf{x}}_\mathrm{g} = \frac{1}{\mathbf{n}} (\log_{10} \mathbf{x}_1 + .. + \log_{10} \mathbf{x}_\mathbf{n}).$$

Any base can be used to compute logarithms for the geometric mean. The geometric mean is the same regardless of which base is used. The only requirement is that the logs and antilogism should be in the same base. For example, these laboratory data are not normally distributed: 1, 10, 100, 1,000, 10,000, and 100,000. The logs of these values are $\log_{10}(1) = 0$, $\log_{10}(10) = 1$, $\log_{10}(100) = 2$, $\log_{10}(1{,}000) = 3$, $\log_{10}(10{,}000) = 4$, and $\log_{10}(100{,}000) = 5$. Then the mean of these values is (0 + 1 + 2 + 3 + 4 + 5)/6 = 2.5. The geometric mean of these data are geomean = $10^{2.5}$ = 316.23.

For symmetric distributions, the mean, median, and mode are the same. However, for highly skewed data, the differences can be considerable. The median in this case is a better representation of the center of the distribution than the mean.

Measures of Spread and Dispersion

The measures of locations are a reflection of the data central tendency, and the measures of variability are a reflection of the data spread or dispersion. The range, percentiles or quantiles, variance, and SD are frequently used statistics for measures of dispersion.

The range is the distance between the lowest and the highest values.

The *percentiles or quartiles* divide the data into parts, for example, the highest third, the middle third, and the lowest third. The 25th and 75th percentiles are also called first and third quartiles, and the median is the 50th percentile.

The *variance* is a measure of the difference of each value from the mean value. Sample variance or variance is defined as follows: if we denote a set of data by $X = (x_1, x_2, \ldots, x_n)$, then the sample mean is typically denoted with a horizontal bar over the variable

$$\overline{\mathbf{x}} = \frac{1}{\mathbf{n}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{x_i}$$

and the sample variance is denoted with

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Variance is an average of the squared distance from each observation to the sample mean.

The *SD* is a measure of dispersion expressed in terms of the original units. A sample SD is the square root of sample variance.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For example, in a group of values of 1, 2, 3, 4, 5, and 6, the mean is 3.5. The SD can be calculated as follows:

$$s = \sqrt{\frac{1}{6-1}((1-3.5)^2 + (2-3.5)^2 + .. + (6-3.5)^2)} = 1.87$$

The variance of these data is 3.5 and the SD is 1.87. For a normal distribution, the mean ± 2 SDs contain approximately 95% of the data.

TYPE OF DISTRIBUTIONS

Most statistical tests are based on the distribution of the data, which describes how often certain values occur, the range of values, and the shape of the probability and value curve. This is necessary to convert ideas and words into probabilities that can be used for statistical decision making. Distributions are based on the probability of the value or range of values of a particular variable. For example, Figure 1 shows the histogram of diastolic blood pressure in a sample of healthy individuals. The horizontal axis is the value of blood pressure in several different ranges, and the vertical axis is the count or number of individuals whose diastolic blood pressure falls within a range (eg, 90 to 95 mm Hg). If we could sample many people and divide the data into smaller and smaller

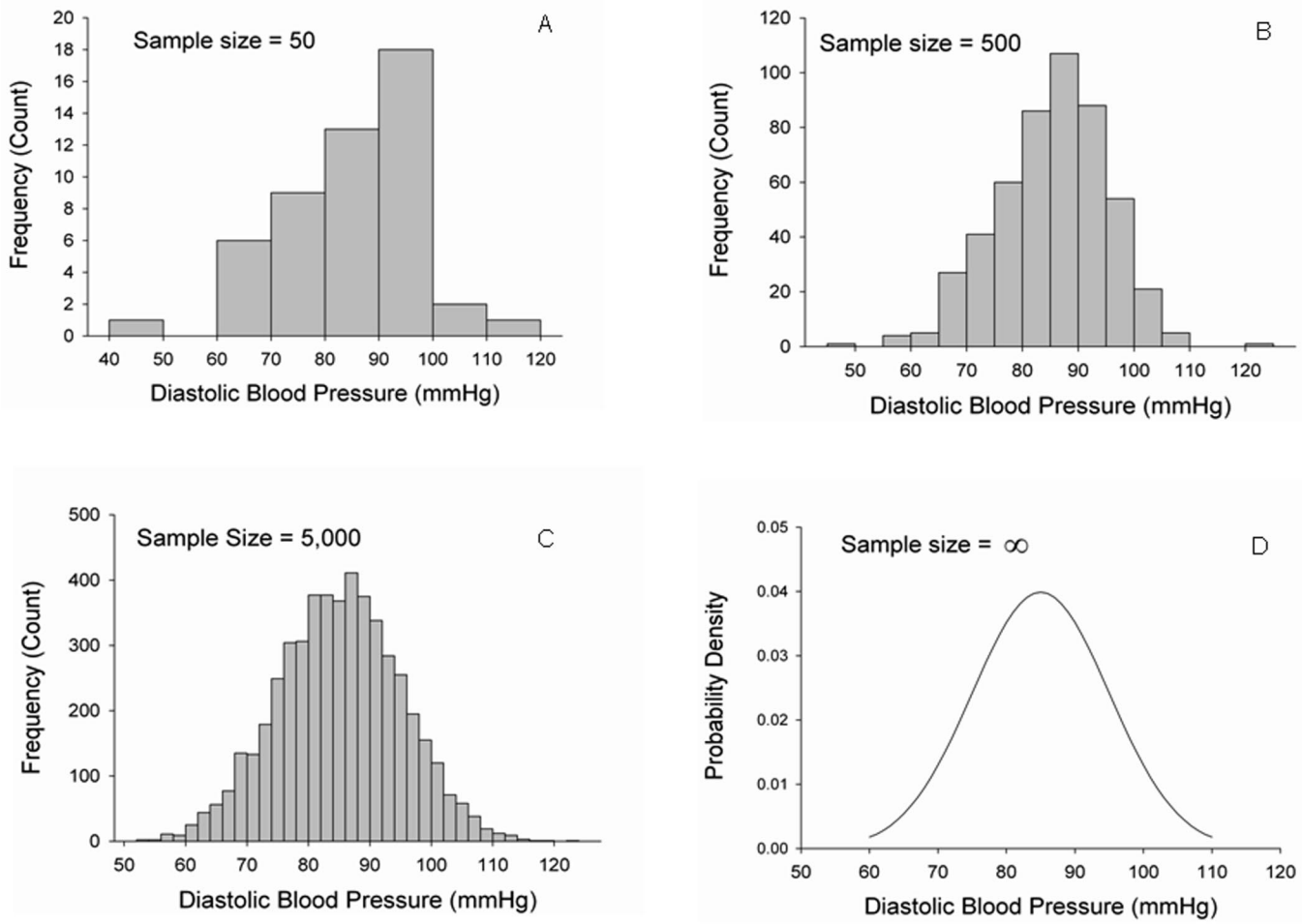


Figure 1. Histograms of the diastolic blood pressure of healthy individuals of different sample sizes. A, Sample size of 50; B, sample size of 500; C, sample size of 5,000; D, sample size of infinity.

ranges, the bumps in the curve would probably smooth out (Fig 1A-D). Finally, if we replaced counts with probabilities (by dividing the height of each bar by the total number of individuals in the sample, we would approximate a statistical distribution, in this case a "normal" distribution. Mathematically, distributions assume sampling an infinitely large number of individuals and classifying them into infinitesimally small ranges. The sum of all the probability densities (the area of a distribution curve) adds to 1.0.

The 4 most commonly used distributions are the normal distribution, Student *t* distribution, binomial distribution, and χ^2 distribution. Their means and variances are described briefly.

Normal Distribution

The *normal distribution*, also called the gaussian distribution, is an important family of continuous probability distributions. Two parameters (location and dispersion) can be used to define each member of the family.

The normal distribution is the most widely used family of distributions in statistics, and many statistical tests are based on the assumption of normality. Fortunately, many variables in nature have a normal distribution; examples include blood pressure, temperature, adult height, spirometry, and many common serum components. The importance of the normal distribution as a model of quantitative phenomena in the biological and behavioral sciences is in part due to the central limit theorem. Many measurements can be approximated by the normal distribution. The normal distribution also arises in many areas of statistics. For example, the sampling distribution of the sample mean is approximately normal, even if the distribution of the population from which the sample is taken is not normal.

The continuous probability density function of the normal distribution is the gaussian function:

$$\mathbf{f(x)} = \frac{1}{\sigma\sqrt{2\pi}}\mathbf{e}^{-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}}$$

where $\sigma > 0$ is the population SD, and the real parameter μ is the expected value (population mean). A normal distribution with mean 0 and variance 1 is called a standard, or unit, normal distribution. This distribution is also called an N(0, 1) distribution.

Figure 1 shows histograms of hypothetical samples of diastolic blood pressure of healthy individuals of different sample size. The frequency count in a particular interval divided by the total sample size can be thought of as probability of occurrence of that interval value (relative frequency). As the sample size (number of individuals samples) increases (Figure 1A-C), the interval gets smaller and the curves become smoother. When the sample size increases to infinity, the curve approaches the symmetric bell shape characteristic of a normal distribution (Fig 1D). The vertical axis (Fig 1D) is in terms of probability density not counts. The area under the whole curve is 1.0, which is true for all statistical distributions. The area under any portion of the distribution curve, for example, between 80 and 90 mm Hg, is the probability of a randomly sampled individual for this population having a diastolic blood pressure between 80 and 90 mm Hg.

The graph of probability density function is plotted in Figure 2 for a standard normal distribution with mean = 0 and SD = 1 and mean = 0 and SD = $\sqrt{2}$. The density function follows a bell-shaped curve, with the mode at the mean and most frequently occurring around the mean. The curve is symmetric around the mean with the point of inflection on each side. The larger the SD, the more spread in the distribution.

Once a set of data are collected, usually the first step is to check its normality. If the normality is met and observation is independent then the method to compute the statistics (ie, arithmetic mean) can be used. Otherwise, the transformation of data are necessary before estimation can be made or special nonparametric methods of analysis can be used.

Several tests can be used to check a given set of data for similarity to the normal distribution. The null hypothesis in this test is that the data set is similar to the normal distribution; therefore, a sufficiently small *P* value indicates nonnormal data. These tests include but are not limited to the Kolmogorov-Smirnov test, Shapiro-Wilk test, and normal probability plot. These tests can be easily performed by using statistical software, such as SAS and SPSS.

The methods for transformation to satisfy the normal distribution include the logarithmic transformation, square root transformation, and reciprocal transformation. Again, once the transformation is made, the normality test is also needed.

Student t *Distribution*

Student *t* distribution and Student *t* tests are 2 different concepts. *Student* t *distribution* (or simply the *t* distribution)

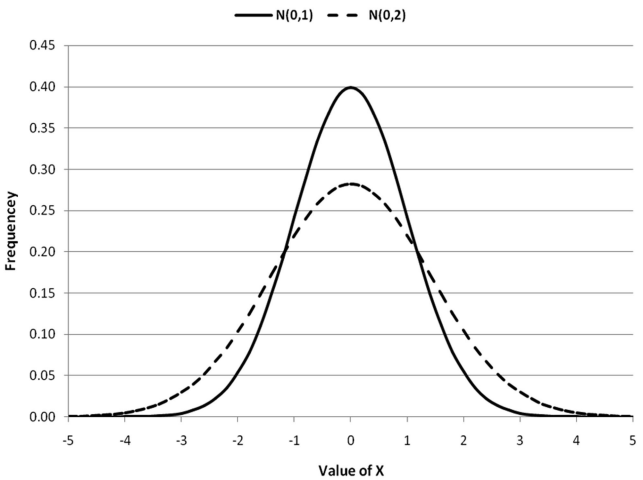


Figure 2. Comparison of normal distribution with mean 0 and variance 1 (N(0, 1)) and normal distribution with mean 0 and variance 2 (N(0, 2)). The horizontal axis is the value of X, and the vertical axis is the frequency at which X occurs (probability).

is a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. Student *t* distribution is the basis of the popular Student *t* test for the statistical significance of the difference between 2 sample means, for confidence intervals, and for the difference between 2 population means.

The Student *t* test is any statistical hypothesis test in which the test statistic has a Student *t* distribution if the null hypothesis is true. It is applied when the population is assumed to be normally distributed but the sample sizes are small.

The overall shape of the probability density function of the *t* distribution resembles the bell shape of a normally distributed variable with a mean of 0 and a variance of 1, except that it is a bit lower and wider. As the number of degrees of freedom increases, the *t* distribution approaches the normal distribution with a mean of 0 and a variance of 1.

Suppose we have a simple random sample of size n drawn from a normal population with mean μ and standard deviation σ. Let $\bar{\mathbf{x}}$ denote the sample mean and s the sample SD. Then the quantity

$$\mathbf{t} = \frac{\bar{\mathbf{x}} - \mu}{\mathbf{s}/\sqrt{\mathbf{n}}}$$

has a t distribution with $n - 1$ degrees of freedom, where s is denoted by

$$\mathbf{s} = \sqrt{\frac{1}{\mathbf{n}-1}\sum_{\mathbf{i}=1}^{\mathbf{n}} (\mathbf{x}_\mathbf{i} - \bar{\mathbf{x}})^2}.$$

The larger the degrees of freedom, the closer the *t* density is to the normal density. This reflects the fact that the SD s approaches σ for large sample size n. The distribution depends on n − 1 but not on the population mean and variance. The *t* test will be discussed in detail in another article in this issue.

The Binomial Distribution

The *binomial distribution* is the discrete probability distribution of the number of successes in a sequence of n independent (success/failure, 0/1, or yes/no) experiments, each of which yields success with probability p. The binomial distribution is the basis for the popular binomial test of statistical significance. A binomial distribution should not be confused with a bimodal distribution. The mean and variance of a binomial distribution (with number of trails n and each trial with a probability of success p) are $\bar{\mathbf{x}} = \text{np}$ and $\sigma^2 = \text{np}(1 - \text{p})$, respectively.

An elementary example is this: roll a standard dice 10 times and count the number of 6's. The outcome in each roll is either a 6 or not a 6. The distribution of this random number is a binomial distribution with n = 10 and p = 1/6. The mean and variance of this binomial distribution are $\bar{\mathbf{x}} = \text{np} = 10*1/6 = 1.67$ and $\sigma^2 = 10*1/6(1 - 1/6) = 1.39$, respectively.

Figure 3 shows the comparison of a binomial distribution with different numbers of trials (n) and each trial with a probability of success (p) of 0.1. As the number of trials increases, the binomial distribution approaches the normal distribution with mean np and variance np(1 − p).

χ^2 Distribution

The *χ^2 distribution* is one of the most widely used theoretical probability distributions in inferential statistics (ie, in statistical significance tests). It is useful because, under reasonable assumptions, easily calculated quantities can be proven to have distributions that approximate to the χ^2 distribution if the null hypothesis is true. The best known situations in which the χ^2 distribution is used are the common χ^2 tests for goodness of fit of an observed distribution to a theoretical one and of the independence of 2 criteria of classification of qualitative data.

If X_i are n independent, normally distributed random variables with mean 0 and variance 1, then the random variable

$$\mathbf{Q} = \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{X}_\mathbf{i}^2$$

is distributed according to the χ^2 distribution with n degrees of freedom. This is usually written as $\mathbf{Q} \sim \chi_\mathbf{n}^2$. In statistics, the phrase *degree of freedom* is used to describe the number of values in the final calculation of a statistic that are free to vary.

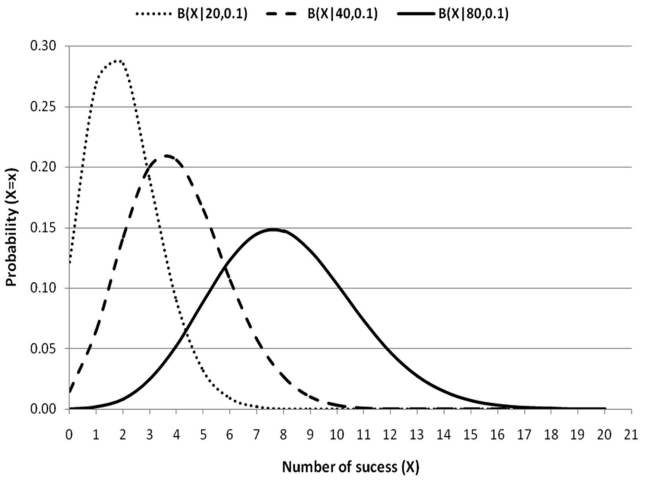


Figure 3. Comparison of binomial distributions with the probability of success at 10% and the number of trials at n = 20, n = 40, and n = 80. Represented by B (X | 20, 0.1), B (X | 40, 0.1), and B (X | 80, 0.1), respectively. With the n increase, the probability of number of successes (X) tends to be normally distributed. The horizontal axis is the number of successes (X), and the vertical axis is the probability that X occurs.

A χ^2 test is any statistical hypothesis test in which the test statistic has a χ^2 distribution when the null hypothesis is true or any in which the probability distribution of the test statistic (assuming the null hypothesis is true) can be made to approximate a χ^2 distribution as closely as desired by making the sample size large enough.

Some examples of χ^2 tests where the χ^2 distribution is only approximately valid are the Pearson χ^2 test, also known as the χ^2 goodness-of-fit test or χ^2 test for independence, and the Yates χ^2 test, also known as Yates correction for continuity. The Mantel-Haenszel χ^2 test, also known as stratified χ^2 test to control for the potential confounders.

A χ^2 test may be applied to a contingency table for testing a null hypothesis of independence of rows and columns. χ^2 tests will be discussed in detail in another article in this issue on statistical tests for more than 2 samples.

SUMMARY

In this article, we introduced basic concepts of statistics, type of distributions, and descriptive statistics. A few examples were also provided. The basic concepts presented herein are only a fraction of the concepts related to descriptive statistics. Also, there are many commonly used distributions not presented herein, such as Poisson distributions for rare events and exponential distributions, F distributions, and logistic distributions. More information can be found in many statistics books and publications.

REFERENCES

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

Requests for reprints should be addressed to:
Runhua Shi, MD, PhD
Department of Medicine
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-4228
E-mail: rshi@lsuhsc.edu

Data presentation

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

INTRODUCTION

Organization and summarization of data are prerequisites for appropriate analysis, presentation, and interpretation. Effective data presentation can be used not only to summarize and communicate findings but also to aid in quality control, data cleaning, and error identification and to help determine the analytical strategy. The presentation process depends on the type of data, the complexity of the information, and the objective of the display.

TYPES OF DATA

Data collected during an experiment or trial can take on a constant value or have varying values. In the case of *constant data*, the observed data has a fixed value (eg, number of eyes or the number of fingers on one hand). Constant data usually form the basis for experimental design rather than observed data. *Variable data* take on varying values that can be classified as qualitative or quantitative.

Categorical Data

Data that can be assigned to one of a set of discrete, usually predefined, categories are termed *categorical data*. In analysis of categorical data, the categories may be used as factors in experimental design. For example, a study may involve the numbers of males and females who have been enrolled in a clinical trial. In this case, *male* and *female* are predefined categories, the numbers of each are the observational data, and the analysis may involve comparison of the number of males vs females. Categorical data may also be treated as observational data and can be used in analyses such as logistical regression. Consider as an example the measurement of B-type natriuretic peptide along with the observation of the heart failure classification of each subject. Logistical regression can be used to define whether there is a linear relationship between B-type natriuretic peptide level and clinical expression of heart failure.

It is not uncommon to create categories from noncategorical data. In place of using age (a quantitative variable) in a pediatric study, one may wish to assign subjects to the categories of neonate, infant, child, and adolescent. Although this may be helpful for clinical purposes, it changes the way this variable can be used in data analysis.

Categorical data are of 3 fundamental types, nominal, ordinal, and interval, based on the relationship among the assigned categories.

Nominal data are categorical data that can be described only by a name, for example, allergy manifestation (rhinitis, asthma, urticaria, or eczema), race or ethnic origin (white, black, Hispanic, or Asian), and state of residence (Alabama, California, Florida, and so on). The categories are not related to each other with respect to value or order. Although their listing does not have to follow a particular order, they may for convenience be arranged alphabetically or according to some other designation, such as frequency of occurrence.

Ordinal (or ranked) data are categorical data that follow a certain order, for example, disease severity (mild, moderate, or severe) or faculty rank (instructor, assistant professor, associate professor, or professor). Ordering does not necessarily imply equal intervals among ranks. The categories may exist as named or numbered categories. For analytical convenience, named categories may be assigned numbers. Numbered categories or numbering schemes, such as 1 for mild, 2 for moderate, and 3 for severe, appropriately designate ranking, but a score of 2 is not necessarily twice the effect as a score of 1, and the difference in disease severity between scores 1 and 2 may not be the same as the difference between scores 2 and 3.

Interval data are categorical data that have a predefined ordering and a constant interval or effect relating the categories. The months of the year represent an interval scale because each is represented by (approximately) 30 days. Assigning quantitative data to 1 of 4 quartiles representing the range of the data, or age to equal bins of 5 years, would also represent interval scales.

Quantitative Data

Quantitative data are data that are numerical and exist over a range of values. The number representing a quantitative observation can be continuous or discrete.

Continuous data are numeric measurements on variables that can take on any value and in which the measurements include a high degree of precision. Data of this type commonly include a fractional component, such as the measurement of cardiac output as 4.36 L/min. Data without fractional components can also be considered continuous if the number of potential values is relatively high, such as B-type natriuretic peptide, which can range from less than 100 ng/L to more than 10,000 ng/L.

Discrete data are those that can take on only certain values rather than any value. These are usually represented as an integral value. Implied in this definition is that the range of values is relatively wide. Although interval (categorical) data

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

Disclosures: Authors have nothing to disclose.

Received for publication January 7, 2009; Received in revised form February 19, 2009; Accepted for publication February 27, 2009.

as described in the previous section could also fit in this definition, interval data are generally limited to a small number of categories, typically less than 5 to 10 categories. In practice, discrete data most commonly represent an integral approximation of a continuous variable, such as age when represented in years or heart rate when represented as beats per minute. Age and heart rate could easily be measured to a much higher precision than their integral values, but the number of age or heart rate values that can be observed is sufficiently large to assume the characteristics of quantitative data. When this is the case, we can apply parametric statistics to these data.

METHODS OF DATA PRESENTATION

Presentation of data may take 3 forms, often combined: text, tables, and graphs. A good rule of thumb, as described by van Belle, is to "use sentence structure (ie, text) for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships."

Text Presentation

Describing the data in a text form is the most common but not always the best method. It is usually the simplest but has a suboptimal readability rate and does not have the visual impact of the other 2 forms. Its comprehension can be enhanced by constructing short sentences and using simple, yet grammatically correct, language.

Tabular Presentation

Tabular presentations are common and popular. Tables can include informative details with a visual effect better than the text's. A table's readability and visual impact depend on its design. Tables can be *simple*, representing 1 variable, or *compound*, representing 2 or more variables. Complex tables with numerous variables have low readability rates.

The principal variable in the table (usually the first column) may be categorized as follows: an *array*, where qualitative variables may be listed in any preferred way, but quantitative variables should be categorized in an ascending or descending order, or a *frequency distribution*, in which distributing the observations in class intervals of a reasonable number to reveal a trend or the type of distribution.

Recommendations on the use of frequency distributions include the following:

- The class intervals should allow the inclusion of every possible observation by having a definite beginning and an end to avoid overlap. For example, a group of children 1 to $<$15 years of age may be classified as 1 to 4 years, 5 to 9 years, and 10 to 14 years; and not 1 to 5 years, 5 to 10 years, and 10 to 15 years. Accordingly, a child who just turned 5 years will be in the second category not the first.
- Avoid open-end intervals (eg, $<$5 years or $>$10 years). Such intervals are more likely to be associated with errors and may result in erroneous computations or graphic presentation.
- Choose a number of categories that can reveal trends and facilitate appropriate analysis. A category that has a low frequency may be combined with another.
- Any categories with low frequencies that cannot be combined with the others may be combined together in one category as "other."
- Minimize the construction of intervals of unequal width. A high frequency may give a wrong impression of importance to a category merely because it is wider than the others. In such situations, adjusted frequencies (according to the width of the category) should be considered. In certain situations, however, the use of unequal intervals of classification would be appropriate, for example, classification according to developmental age: infancy (0 to $<$1 year), childhood (1 to $<$12 years), adolescence (12 to $<$18 years), and adulthood ($\geq$18 years).
- The frequencies may be expressed as absolute numbers, percentages (relative frequencies), or preferably both. The inclusion of actual numbers is particularly necessary when the frequency is low.
- In presenting computed values, rounding the number and choosing the place of the decimal point follow common sense (ie, according to the desired degree of precision or the relative value of the fraction compared with the whole number).
- Cumulative frequencies can be categorized in an ascending or descending manner. The former starts with a "less than" category that includes the lowest frequency, whereas the descending cumulative frequency starts with a "greater than" category that includes the total number.

Table Composition

The informative value of the table depends on its construction. The following principles should be helpful:

- The title should be concise yet sufficiently descriptive to stand by itself without requiring the reader to search the text for clarifications. Many prefer reading tables rather than text.
- The rows and columns should have clear headings and subheadings (if applicable) and must include the relevant unit of measure (eg, number, percentage, lb, kg, U/mL, mg/dL; the latter should not be expressed as mg%).
- A footnote below the table should be for clarifications or spelling of abbreviations that have been indicated by specific symbols in the table.
- If the table includes data from previous publications, the source should be included as a footnote.
- Tables being prepared for publication need to follow instructions provided by the publisher.

Graphic Presentation

Graphs, also referred to as figures, charts, or diagrams, have the quickest, strongest, and most long-lasting visual impact but often lack precision. Their impact can be enhanced by inserting selected data (eg, the number of subjects in each group). Graphs are excellent for demonstrating trends or

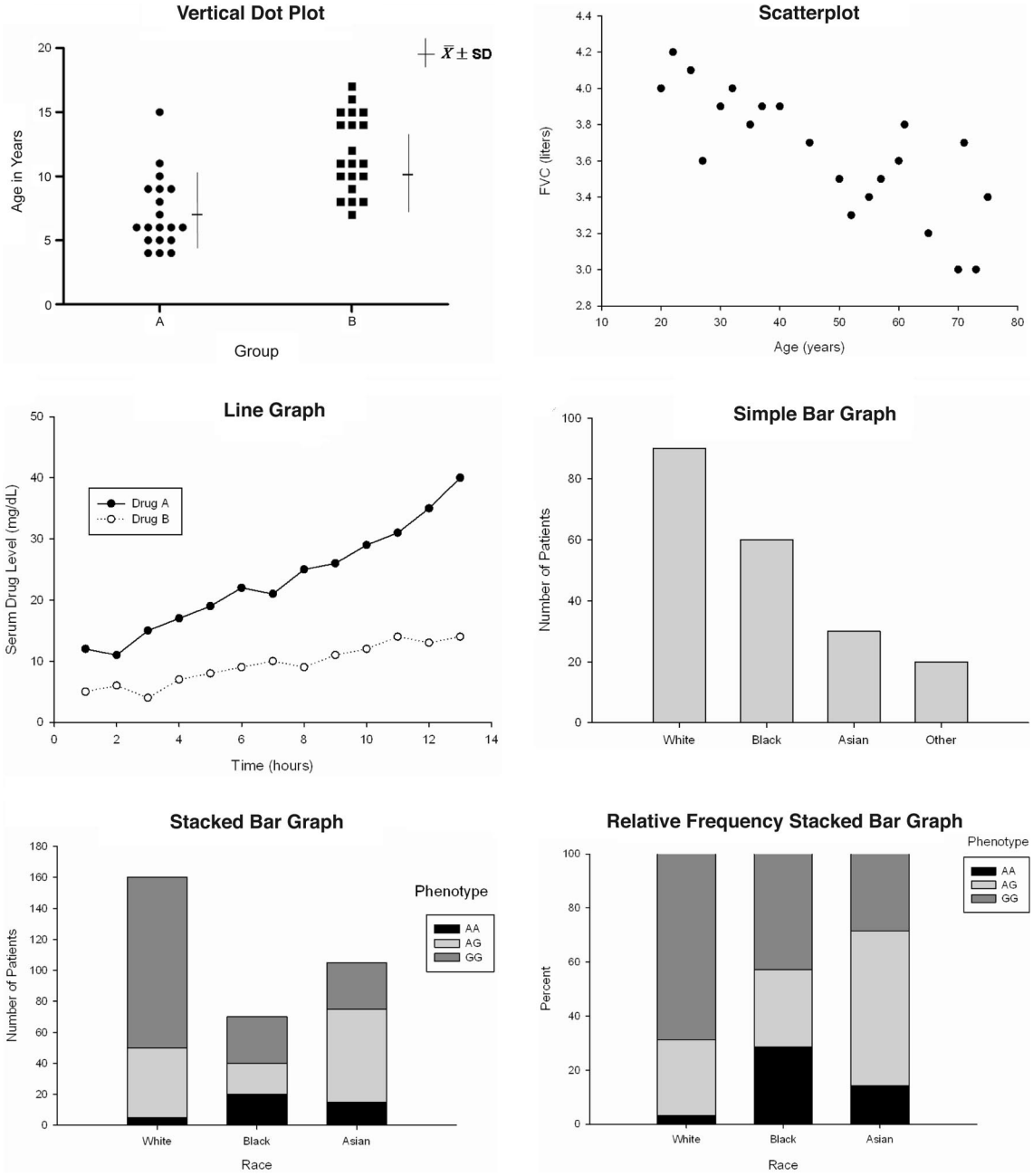
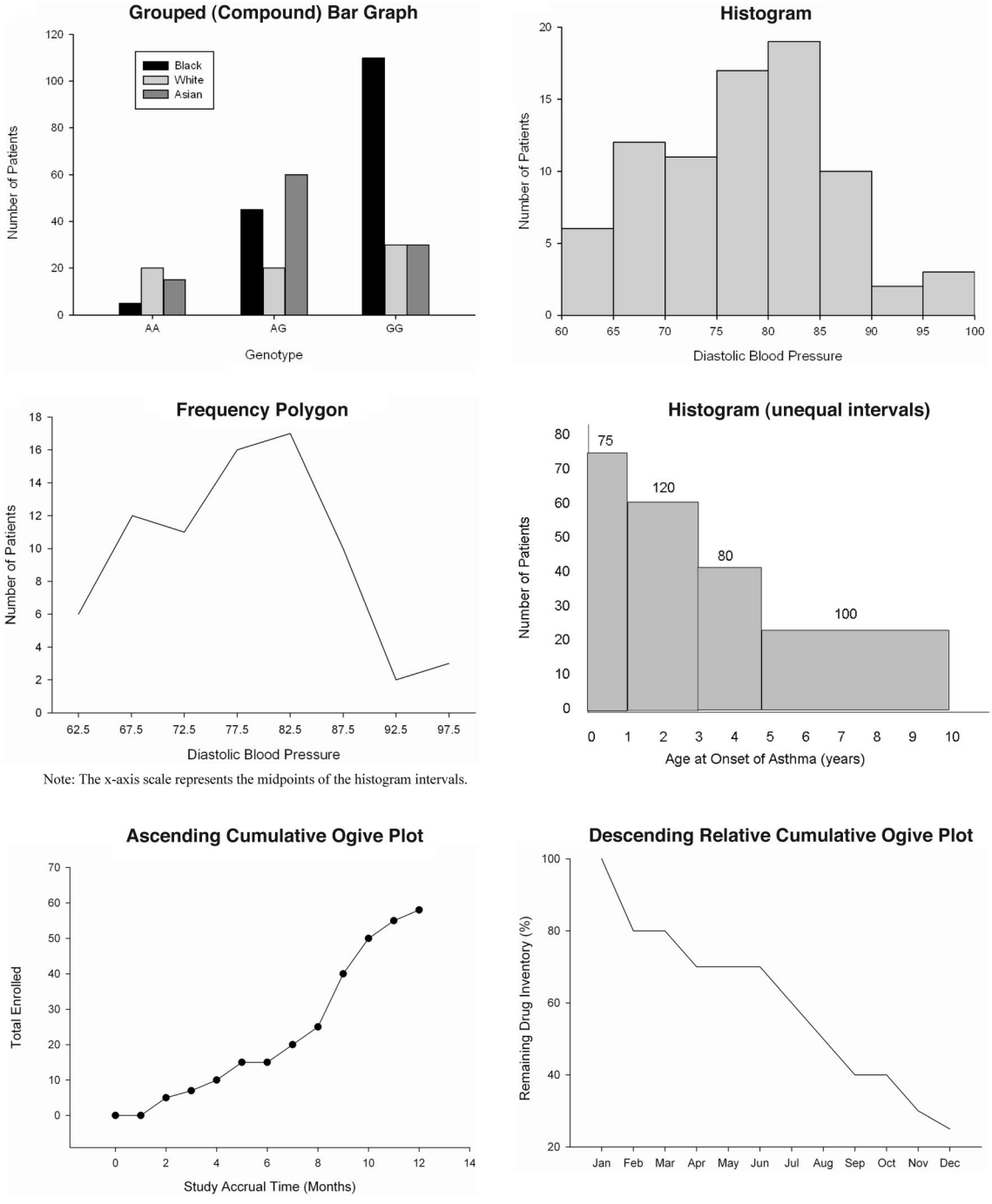


Figure 1. Examples of common graphic presentations.

comparisons. Information displayed in a graph is often derived from tabulated data. Graphs may be used as a substitute for tables or in addition to tables to highlight certain data selected from a table and are especially good when relationships among variables are complex.

Most graphs require a vertical scale (*y*-axis) and a horizontal scale (*x*-axis); one may represent numerical values and the other may represent quantitative or qualitative variables. Equal intervals on an arithmetic scale should represent equal differences in value. The graph's title should be sufficiently descriptive yet concise. Whereas a table's title is placed at the top, the graph's title has been traditionally placed below the graph, at least in publications.



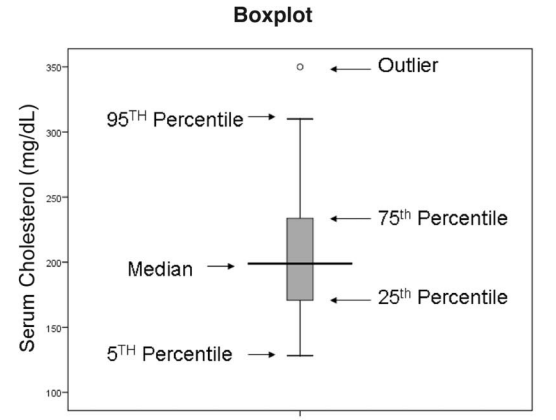
Note: The x-axis scale represents the midpoints of the histogram intervals.

Fig. 1—Continued

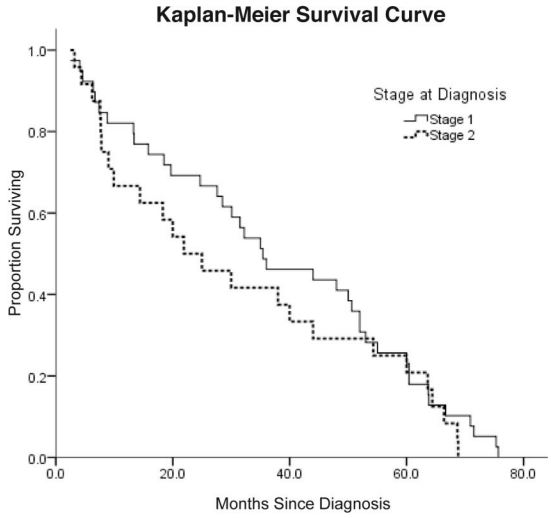
Types of Graphs

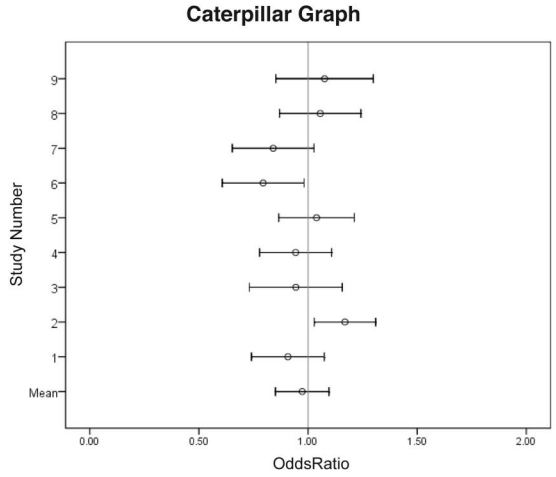
The advances in computer technology have markedly facilitated the production of a large variety of graphic presentations, including complex ones. Several software packages are commercially available and continuously updated. The most common basic types are briefly discussed herein. Examples of common graphs are presented in Figure 1.

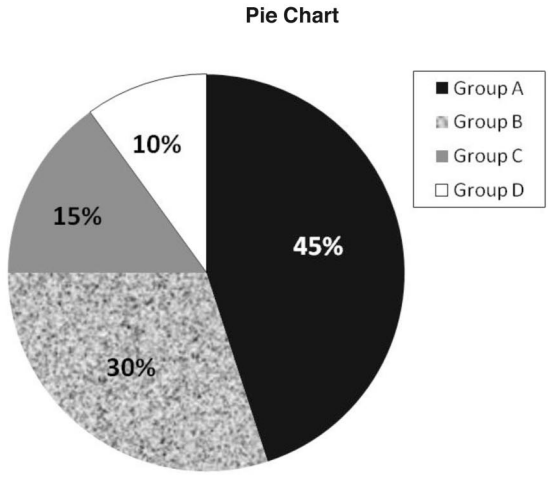
The *dot graph* has the advantage of individual representation of each observation in a linear fashion and thus shows the range, pattern of spread, and any outlying observations. The

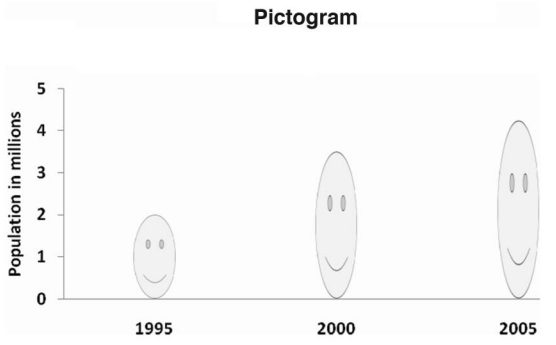


Note that there are several methods of drawing the "whiskers", the vertical lines above and below the box. Some use the interquartile range (IQR) , 75th percentile - 25th percentile, and the first point 1.5 IQR above (and below) the box as the ends of the whiskers. Outliers are more than 1.5 IQR from the box.









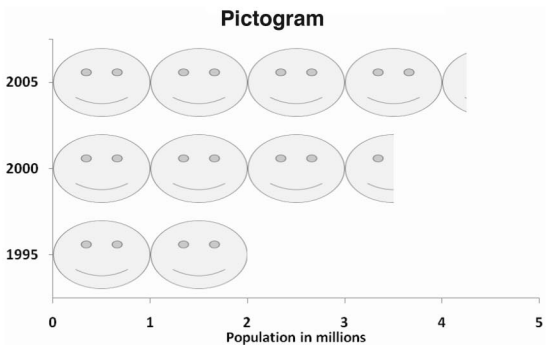


Fig. 1—Continued

mean may be added as a horizontal short line and the standard deviation (SD) or standard error (SE) by vertical lines. Alternative to the mean, the median and the lower quartile (25th percentile) and upper quartile (75th percentile) can be displayed. A dot plot is suitable for a small number of observations and can be used for presentation of multiple groups. Too many observations may produce a clutter and overlapping of too many dots.

The *scatter diagram* is a dot graph where each dot represents 2 measurements for each individual observation (eg,

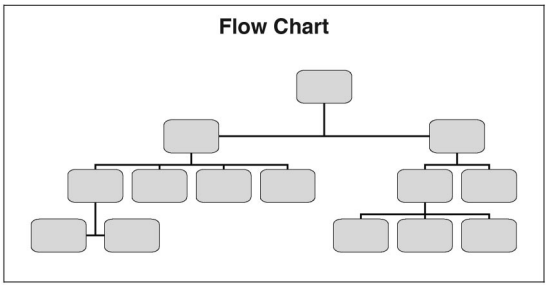


Fig. 1—Continued

height and weight). Hence, it gives a picture of correlation. A regression line or idealized curve can be superimposed on a scatter diagram.

The *line graph* is good for revealing a relationship or trend (ie, the change of numeric variable on the vertical axis in relation to an ordinal or a quantitative variable on the horizontal axis). It can be a single line or multiple, each representing a separate set of data. The line graph can allow presentation of the trend of 2 variables that have different units. In this case, the first variable is assigned the left vertical axis and the second variable is assigned the right vertical axis.

The *bar graph* is commonly used for qualitative or discrete quantitative variables. The height of the bar represents the value and in case of a mean may be combined with a vertical line representing the SD or SE. All tables and graphs should specify whether SD or SE are shown. It is a common mistake not to make this clear. The bars must be of the same width, and the distance in between can be half or equal to the bar's width. In a complex bar graph where multiple measurements are being represented, the bars of each set can be next to each other without spacing. In certain situations, the bar can be divided into parts representing the share of the bar's components, whether as absolute numbers or percentages. If the bar labels are lengthy, the axes may be switched.

The *histogram* is actually a set of bars next to each other without spacing to represent a continuous quantitative variable. The horizontal axis represents the numerical variable and the vertical axis represents the frequency (absolute or relative). It differs from the bar graph in that the width of the bars may vary and each should reflect the width of the respective category and that the vertical axis represents counts of values not values themselves. Hence, the area, not the height, of the bar represents the frequency in that category. Histograms are useful to identify extreme values and to identify the underlying distribution of the measures (eg, normal, log normal), which are important to the choice of statistical analytical technique.

The *frequency polygon* is practically a line graph connecting the midpoints of the tops of the bars of a histogram. Hence, it can be used for multiple sets of data without plotting the histograms. It can also demonstrate the shape of distribution (eg, normality, skewness, or kurtosis) as addressed in detail in the preceding article on descriptive statistics.

The *box graph* (or box plot) is useful in presenting a measure of central tendency together with a certain spread of dispersion, for example, the mean with spreads of ± 1 SD and/or ± 2 SDs or the median (50th percentile) with spreads of certain percentiles (eg, 25th and 75th and/or 10th and 90th). The 25th and 75th percentiles are also called first (lower) and third (upper) quartiles, respectively, and the length of the box is called the interquartile range. Most commonly, medians and interquartile ranges are used for box plots. The figure annotation should specify what the symbols mean. It has the advantage of excluding any outlying or extreme observations (eg, below the first quartile or above the fourth quartile). Box plots also indicate presence or lack of symmetry in the distribution.

An *ogive* is a line graph representing cumulative frequencies, either ascending (begins with the lowest category) or descending (begins with the total of all categories).

Caterpillar plots are being increasingly used with the surge of meta-analysis studies. The data are gathered from previously published trials selected according to certain criteria to

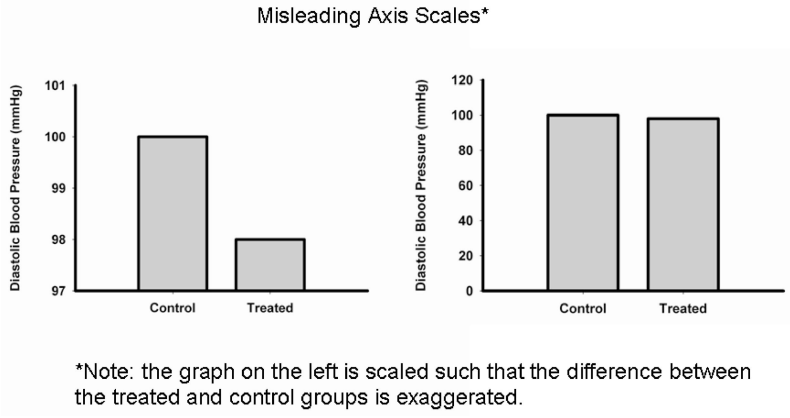


Figure 2. Example demonstrating the misleading visual effect of using an inappropriate scale.

maximize the homogeneity and soundness of the methods. It is a systematic review by experts who survey the literature to identify and characterize the quality of data to be included then calculate odds ratios (or relative risks) with a 95% confidence interval (CI). In addition, an overall odds ratio (displayed at the bottom of the graph) is estimated, taking into consideration a "weight" assigned to each study.

In the caterpillar chart, the conclusion about the findings of each trial is determined by the location of its respective odds ratio together with its whole 95% CI in relation to a line representing an odds ratio of 1.0, with a decreasing scale to the left and an increasing scale to the right. Overriding indicates nonsignificance.

Survivorship or *actuarial graphs* represent changes or probability of occurrence of an event over time relative to the starting point. A well-known example is the Kaplan-Meier graph, which is frequently used to demonstrate the natural course of a disease outcome or to compare different interventional methods on the long-term outcome.

A *pie graph* shows the relative distribution of a total circle into sections (slices of a pie). The magnitude of each component is commonly displayed numerically as a percentage (plus absolute numbers) inside or next to each slice of the circle. Multiple pie charts can be used to compare different sets of data.

Pictograms represent the variable as a picture the size of which reflects its approximate magnitude in contrast to the size of another similar picture. Another method can be representation by a number of pictures of a constant size. Each figurine represents a fixed amount, and a part of a figurine represents a fraction of that amount. Pictograms markedly lack precision and are usually used for commercial advertisements.

A *flow diagram* is used to demonstrate the stages of a study's method or outcome, either in a vertical or a horizontal direction and often combined.

Graph Composition

The type of data being presented and the main objective of the presentation would determine the most appropriate choice of the following:

- Type of graph.
- The measurement scale regarding spacing.
- The presence of a scale break. If a scale break is used, the 2 points on both sides of the break should not be connected.
- The starting point of the scale (ie, zero or another value).
- The choice of an arithmetic or logarithmic scale. Many biological variables follow logarithmic or exponential distributions and may best be represented on a logarithmic scale.
- Clear labeling of the scales, including identification and units (eg, number, percentage, kg, mg/dL).
- A legend may be needed to supplement the labeling and is best placed in the right upper quarter of the graph.
- A reference line can be included horizontally across the entire graph. A reference range can be included as 2 lines or as a lightly shaded area that does not hide the data.
- The choice of scale is an important issue. Small differences can be made to appear large by expanding the scale of the vertical axis to encompass a small range of values. Conversely, important differences can be masked by compressing the scale to cover too large a range (Figure 2). Where practical, it is good to have a scale starting with the value 0 or at least the typical range of numbers that could be observed. It is also important with a series of graphs to make the scales cover a similar range so that the graphs are visually comparable.

REFERENCES

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

Requests for reprints should be addressed to:
Sami L. Bahna, MD, DrPH
Department of Pediatrics
Allergy and Immunology Section
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-3932
E-mail: SBAHNA@LSUHSC.EDU

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.

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.

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 in 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 chose 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

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 chose 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: $\mu_A = \mu_B$ 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 if the null hypothesis were true. At degrees of freedom of 98 (n1 + n2 − 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, 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 we decided that 10 mm Hg or greater would be of more clinical significance, then only 16 patients would be needed

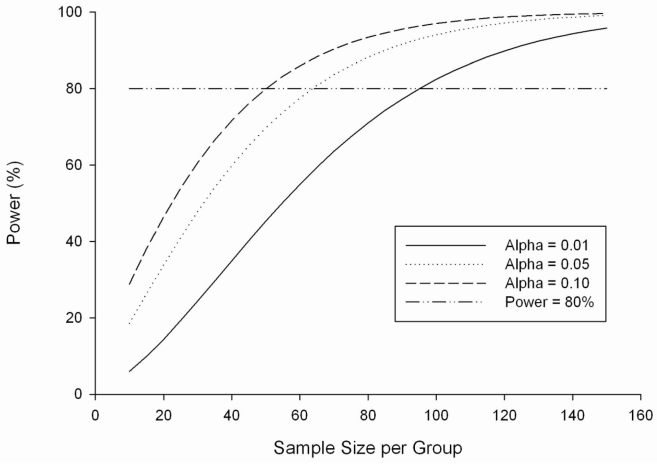


Figure 1. The effect of type I error (α) on the required sample size.

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 impressive. 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

Statistical tests for 1 or 2 samples

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

INTRODUCTION

The use of statistics can be categorized as either descriptive or inferential. With inferential statistics generalizations are made about populations from samples. For example, how representative are the values calculated from a sample, how likely is the result a chance finding, and how do different groups compare? As described in the previous article by McLarty and Bahna, the basics of hypothesis testing are as follows: make a hypothesis, draw a sample, calculate a test statistic, and, based on how probable the test statistic is, decide whether to reject the hypothesis. The tests described herein are among the most commonly used tests for hypothesis testing. Each procedure is described for a single-group test and for 2-group tests. They are divided into parametric and nonparametric tests of hypotheses.

Parametric tests are tests based on known statistical distributions (eg, normal distribution, Student distribution). The name is derived from the fact that these distributions have parameters in the equations that describe them (such as the mean and standard deviation (SD) for the normal distribution). Ideally, parametric tests should only be used on data that are known to be based on a distribution or closely follow the chosen distribution. In contrast, nonparametric tests do not make any assumptions about the underlying distribution and thus are also known as *distribution-free* tests. This topic is explored further later in the article.

TESTING OF MEANS (STUDENT *t* TEST)

Perhaps the most well known of statistical tests are the *t* test and the χ^2 test. The *t* test is used to compare 2 means from independent normal distributions and the χ^2 test to compare grouped data with no explicit distributional assumptions. However, the *t* test can also be used for a single-group comparison, comparing a mean value from a normal distribution to a single-group value. Other single-group tests include testing proportions or percentages from binomial distributions (for example, yes/no data) or from Poisson distributions (counts of events).

ONE-SAMPLE *t* TEST

The Student *t* test is used to compare 2 means, that is, means in 2 different groups (say, treated or untreated), or to compare a single mean against a known constant (eg, normal temperature, average IQ). The idea behind a *t* test is that, under certain conditions, the mean of a sample divided by its standard error (SE) follows a known distribution, the *t* distribution, from which probabilities can be calculated and statistical hypotheses can be tested.

For example, assume a study was performed to test whether the average normal body temperature was really 98.6°F, as generally believed. In this study, assume temperatures were measured in a group (ie, one sample) of 100 healthy persons throughout the day: the mean temperature was found to be 98.1°F and the SE ($SE = SD/\sqrt{n}$) was 0.05. A 1-sample *t* test could help answer the question of whether this finding was statistically different from 98.6°F:

The null hypothesis is that there is no difference, that is, mean of the sample is 98.6 or H_0: $\bar{X} - 98.6 = 0$ (the mean of the test sample is denoted by $\bar{X}$).

The formula for the test statistic in this case is

$$t = \frac{\bar{X} - C}{SE(\bar{X} - C)} = \frac{\bar{X} - C}{SE(\bar{X})},$$

where C is a constant (the SE of a mean minus a constant is the SE of the mean because a constant has no variability). For the example given,

$$t = \frac{\bar{X} - 98.6}{SE(\bar{X} - 98.6)} = \frac{98.1 - 98.6}{0.05} = -10.0$$

The *t* distribution is symmetric so positive and negative test statistics have the same probability; the sign of the statistic can be ignored in computing *P* values.

The *P* value (the probability of a *t* statistic of this magnitude or greater if the means actually were the same) is quite small, $P < .001$. So the conclusion is that, for this sample at least, the average temperature is significantly lower than 98.6°F.

The shape of the *t* distribution, and the calculation of *P* values, depends on an often misunderstood number called the degrees of freedom (df). In the 1-sample *t* test, if n subjects are in a sample, the degree of freedom is n − 1 (because 1 *df* was used up in the calculation of the mean value).

TWO-SAMPLE *t* TEST

The same principles apply for a 2-sample *t* test, and all its variants, as for the 1-sample test: a *t* statistic is the ratio of a mean value divided by its SE. The complication comes in estimation of the SE. The specific formula depends on whether the 2 groups being compared have the same variance and whether the sample sizes are the same in both groups. In a 2-sample test the numerator of the ratio is the absolute difference between 2 means, $\bar{X}_1 - \bar{X}_2$ (ie, whether it is negative

Affiliations: *Department of Medicine, Cancer Prevention and Control, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana; †Departments 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 February 10, 2009; Received in revised form March 27, 2009; Accepted for publication March 30, 2009.

or positive). The denominator is the SE of the difference, that is, $SE(\bar{X}_1 - \bar{X}_2) = \sqrt{SE(\bar{X}_1)^2 + SE(\bar{X}_2)^2}$. So, the test statistic in this case is given by the following:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SE(\bar{X}_1)^2 + SE(\bar{X}_2)^2}}$$

It can be easily seen that if 1 of the samples is replaced by a single number, then this formula is equivalent to a 1-sample *t* test.

Note that the degrees of freedom for a 2-sample *t* test is $n_1 + n_2 - 2$ (2 *df* are used in calculating the 2 means), where n_1 and n_2 are the sample sizes for each of the 2 groups.

A hypothetical example is in testing a potential therapy for asthma control, in which a group of 50 patients were given the therapy and compared with a group of 40 patients with similar mean forced expiratory volume in 1 second (FEV_1) given a placebo. The mean percent predicted FEV_1 was found to be 80% (SE, 2.0%) in the treated group and 76% (SE, 2.5%) in the untreated group. A Student *t* test to see if the treatment made a difference yields the following:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SE(\bar{X}_1)^2 + SE(\bar{X}_2)^2}} = \frac{80 - 76}{\sqrt{2^2 + 2.5^2}} = 1.25$$

The degrees of freedom is $50 + 40 - 2 = 88$; from standard statistical tables or statistical software, it can be seen that the corresponding *P* value was .11. Because $P > .05$, the null hypothesis of no difference in means cannot be rejected. We conclude that the treatment was not significantly effective.

Note that the *t* tests described herein work well if the variances of the 2 compared means are roughly equal. If they are not, then the underlying distribution is no longer a *t* distribution. Various correction methods have been devised to get around this; one of the most commonly used methods is the Welch *t* test, which uses a more complex calculation of degrees of freedom and allows an approximation to a *t* distribution. The question remains, "How much difference in variances is allowed?" One solution is to always assume unequal variances. If the variances are equal, the *t* calculations and *P* value calculations simplify to the ones discussed herein. The unequal variance calculations are cumbersome but with software readily available and even free online sites that will do the *t* test calculations, it is easy to use the unequal variances versions.

PAIRED *t* TEST

All statistical procedures have underlying assumptions. For the *t* test, the major assumptions are that the samples are taken from a normally distributed population and the groups being compared are independent. However, in many medical studies, groups are not independent. The most common example is before-after studies, where something is measured in the same individuals before and after an intervention of some kind. In this case, the before and after data are not independent because the same individuals are included in both sets of measures. Other examples include matched studies in which individuals in each group are deliberately matched closely with individuals in the other group. Twin studies are another example of nonindependent groups in which the outcome measure is available on each twin.

Fortunately, there is an easy way around this problem. If the difference in outcome measures is computed for each pair of persons, then a 1-sample *t* test can be performed on the differences:

$$t = \frac{\bar{X}_D - C}{SE(\bar{X}_D)} = \frac{\bar{X}_D - 0}{SE(\bar{X}_D)} = \frac{\bar{X}_D}{SE(\bar{X}_D)}$$

where $\bar{X}_D$ is the mean of the differences.

The null hypothesis in the paired test is that the mean of the paired differences is 0. The degrees of freedom for a paired *t* test is the number of pairs − 1.

In the following hypothetical example, a cholesterol-lowering drug is given to patients. The cholesterol level is measured before treatment and after 6 months of treatment. A paired *t* test is used to see if the treatment was effective.

Patient No.	Before treatment	After treatment	Before-after difference
1	290	210	80
2	300	310	−10
3	250	210	40
4	220	215	5
5	195	185	10
6	230	235	−5
7	245	150	95
8	265	265	0
9	275	250	25
10	260	270	−10
Total	–	–	230
Mean value	–	–	23
SE	–	–	11.9

The null hypothesis is that the mean of the before-after differences is zero. In this example, it is 23.

$t = \bar{X}_D/SE_D = 23/11.9 = 1.93$. The degrees of freedom are 9 ($P = .085$). We cannot reject the null hypothesis, so we conclude the treatment was not significantly effective.

In this example it would be an error to compare mean cholesterol levels before treatment with mean cholesterol levels after treatment with a 2-sample *t* test: the measurements did not come from 2 independent groups but the same group measured twice. As discussed, there is an underlying assumption of normality in *t* test data. Although this is the heart of the theoretical basis for *t* tests, *t* tests are relatively insensitive to this assumption, especially if the sample sizes are large (say $n > 30$).

TESTS OF PROPORTION: CATEGORICAL DATA

Simple Proportions

Sometimes it is desirable with a single sample of patients to determine what proportion of them have a certain condition

(eg, atopy) and to determine whether this proportion is unusually higher or lower than expected. There are several ways to go about this statistically, including binomial, multinomial, and Poisson tests. One such test presented herein is the χ^2 test. For binary data (eg, yes/no data) there are better methods, but the single-group χ^2 method leads didactically into the next section of comparing 2 groups with categorical data.

As a simple example, a coin is flipped 20 times and 15 times it comes up heads, that is, 75% of the time. Is this a fair coin? We would expect approximately 10 heads and 10 tails, that is, heads 50% of the time. The χ^2 test compares observed values (15 heads, 5 tails) with expected values (10 each). A test statistic can be derived from the observed and expected values and a *P* value calculated.

The test statistic is as follows:

$$\chi^2 = \sum \frac{(\textit{observed value} - \textit{expected value})^2}{\textit{expected value}} = \sum \frac{(O - E)^2}{E}$$

In words, the sum of the squared differences between observed and expected values divided by the expected value.

For example, $\chi^2 = (15 - 10)^2/10 + (5 - 10)^2/10 = 5.0$ and $P = .025$. We can safely conclude that this coin is biased: if the coin were fair we would rarely (1 time of 40) expect to have a value of $\chi^2 \geq 5$. Calculation of *P* (or looking it up in a table) depends on the degrees of freedom: if there are 2 categories in our sample (heads/tails), the degrees of freedom would be $2 - 1 = 1$. If there are 3 categories, the degrees of freedom would be $3 - 1 = 2$.

Note that in this case it is obvious what the expected values should be, half heads and half tails. The expected values will vary with the situation. For example, if it is known that 30% of the population is clinically obese, a local sample of individuals could be compared against this percentage: expected values of 30% obese and 70% nonobese. Most often, perhaps, expected values are assumed to be equally distributed among the categories.

The χ^2 test can be used even when multiple categories are possible. Gene frequencies, for example, are often expressed as percentage normal (wild type), percentage heterozygous for a particular mutation, and percentage homozygous alleles. A simpler example is rolling a single dice, where a fair die would have an equal chance (1 of 6) for each of the 6 possible numbers to come up. In this case, the degrees of freedom would be $6 - 1 = 5$.

More Than 1 Group

The advantage of the χ^2 test is its simplicity; the same schema of comparing observed to expected values will work for 2 or more groups and for multiple possible outcomes. In principle, data that can be put into a table can be analyzed with a χ^2 test, provided certain constraints (discussed herein) are met. An example follows.

Assume we want to compare the gene frequencies of a gene mutation possibly involved in an allergic disease for 2 different racial groups, Asians vs whites.

Subjects	AA	AG	GG	Total No. of Subjects
Asian	15 (7.9)	60 (41.6)	30 (55.5)	105
White	5 (12.1)	45 (63.4)	110 (84.5)	160
Total	20	105	140	265

The numbers in parentheses are expected values. For tables like this, the expected values for a given cell (say the AA genotype in Asians) is column total (20) times row total (105) divided by the overall total (265). This assumes that the column values are distributed proportionately in each row, that is, that there are no racial differences in allele frequencies. In this example it is easy to see that most Asians (60 of 105) have the AG alleles, whereas most whites (110) have GG. However, it is not always so obvious, and in any case it is still desirable to see whether this pattern is statistically significant. The calculations are as follows:

$$\chi^2 = \frac{(15 - 7.9)^2}{7.9} + \frac{(60 - 41.6)^2}{41.6} + \frac{(30 - 55.5)^2}{55.5} + \frac{(5 - 12.1)^2}{12.1} + \frac{(45 - 63.4)^2}{63.4} + \frac{(110 - 84.5)^2}{84.5} = 43.3$$

For tabular data the degrees of freedom is number of rows minus 1 multiplied by the number of columns minus 1, or in this example $(2 - 1) \times (3 - 1) = 2$. From tables or a computer program it can be seen that the *P* value (probability of type I error) is very small, $<.0001$. The samples from each group were random, the racial data are independent, and none of the expected values are less than 5; the χ^2 assumptions are met. So, it is extremely improbable that the observed racial difference in genotypes is due to chance alone.

Limitations of the χ^2 Test

Like all statistical tests, the χ^2 test has its assumptions and limitations. The major assumption is independence of the groups being compared. An important limitation is that the formula involving observed and expected values is only an approximation to the χ^2 distribution. The approximation breaks down with sparse tables (tables with small numbers in the cells). Two rules of thumb are that no expected value should be less than 1.0 and no more than 20% of the cells should have expected values less than 5.0. In general, the χ^2 approximation is good if neither of these 2 rules is violated. What should be done if 1 or both of these rules are violated? One solution is to collapse (combine) 2 or more rows and/or columns so that each cell has more numbers of subjects. Another possibility is to use a statistical method that does not use an approximation to a distribution: Fisher's exact test. Fisher's exact test is beyond the scope of this discussion but is readily available in most statistical programs for small tables. It does not have to follow the 2 rules mentioned herein. Fisher's exact tests become computationally intensive for larger tables, and special methods have been developed to compensate for this problem.

NONPARAMETRIC TESTS

With the exception of the χ^2 test, all of the tests of hypotheses discussed herein require assumptions about the distribution of the data samples, for example, the *t* distribution. Means and standard errors are *parameters* that help specify the underlying distribution. Therefore, *t* tests and others like it are called *parametric tests*. Fortunately, many variables in nature follow a normal distribution. However, there are situations in which the underlying distribution is not normal or is not known or is too complex for practical use. Or, perhaps, the data are not numeric at all (eg, Likert scales), such as strongly disagree, disagree, no opinion, agree strongly, or disagree. For these situations a body of statistical tests called *nonparametric* tests are used; these are tests that do not require knowledge of distribution parameters. The simplest example of a nonparametric test is the χ^2 test. The underlying assumption is that groups being compared are independent (eg, not from the same population or closely related subjects or matched). The null hypothesis is that the rows and columns of the data table are distributed proportionately.

One-Sample Nonparametric Tests

For most common statistical tests, such as *t* tests, that are based on normally distributed data there is a nonparametric equivalent. The 1-sample χ^2 test discussed herein is a nonparametric test. It makes no assumptions about the underlying distribution of the data. It can also easily be extended to 2 or more groups. Another 1-sample nonparametric test is the Kolmogorov-Smirnov *z* test, which is often used to test whether data come from a normal distribution. This will not be discussed herein. However, there is another important class of nonparametric tests based not on raw data values but ranks of data (sorted lists of data). Intuitively, it is reasonable to think that if the groups had unequal raw data values (one group higher than the other, for example) their average ranks or sum of ranks would also be unequal. This is the basis for many nonparametric tests. This will be illustrated with the 2-sample discussion herein.

Two-Sample Nonparametric Tests

Consider a hypothetical example of test scores in 2 groups, boys vs girls. The raw data may look like the following table:

t scores	
Boys	**Girls**
70	60
90	50
85	95
55	80
65	75

A 2-sample *t* test might be used to compare the mean scores between boys (73) and girls (72); however, from the small sample given, we may have no assurance that the data are normally distributed. A nonparametric test should be considered. The data could be ranked (sorted) without regard to sex yielding the following:

Rank	Score	Sex
1	50	Girl
2	55	Boy
3	60	Girl
4	65	Boy
5	70	Boy
6	75	Girl
7	80	Girl
8	85	Boy
9	90	Boy
10	95	Girl

Replacing the raw data in the original with the ranks and adding the ranks in each column yields the following:

Boys	Girls
2	1
4	3
5	6
8	7
9	10
Sum = 28	Sum = 27

Surprisingly, no matter how the original data are distributed, the sums of ranks do have predictable distributions for which probabilities (*P* values) can be calculated. In the simple example given, the *P* value is .92, indicating that test scores are not likely to be different between the sexes. The null hypothesis of no difference could not be rejected.

This test is called the Mann-Whitney test or sometimes the Wilcoxon rank sum test. When the raw data have ties between the groups, an average rank is assigned to both groups for the tied data. Although the computations for ranks and sums of rank are not particularly onerous, *P* values still have to be determined from elaborate tables or computer programs. Most statistical packages contain the Mann-Whitney test. Note that it is not required to have the same number of subjects in each group.

Paired Nonparametric Tests

Just as for the paired *t* test, there is a nonparametric paired test for comparing groups that are not independent, for example, twins, before and after measures on the same persons, and closely matched groups. With the paired test, the difference of each pair is computed. If the groups were equal then you would expect that there would be a balance of negative and positive values. If the differences are ranked, as in this case, you would expect equal sums of positive and negative ranks. The Wilcoxon signed rank test uses this idea of summing positive and negative ranks to come up with a test statistic. The example below should illustrate this method.

In a study conducted by 1 of the coauthors (J.W.M.), 25 men with prostate cancer were given a drug to lower prostate specific antigen (PSA). Results from the first 10 men are given in the table below. PSA levels were measured on each man before the administration of the drug and 6 weeks later. Each man was essentially his own control. Differences were calculated for each pair; the differences were ranked without regard to sign, and then the original signs affixed to the ranks. Finally, the positive ranks and negative ranks are summed separately and a *P* value computed using the smaller of the 2 sums.

Patient No.	PSA before	PSA after	Difference (before − after)	Rank (disregarding signs)	Signed rank
1	10.2	9.4	0.8	3	3
2	19.1	10.5	8.6	10	10
3	9.5	5.6	3.9	7	7
4	9.9	11.3	−1.4	4	−4
5	10.2	10.3	−0.1	1	−1
6	9.3	5.2	4.1	8	8
7	7.1	5.3	1.8	5	5
8	20.6	15.8	4.8	9	9
9	15.5	13.4	2.1	6	6
10	10.2	9.4	0.7	2	2

The sum of all the negative ranks in the last column is $W_- = -5$.

The sum of all the positive ranks in the last column is $W_+ = 50$. From tables, using 5 as the test statistic (the smallest of the 2 sums in absolute value) or from a computer program it can be shown that $P = .022$. This is less than .05, so we can reject the null hypothesis of no drug effect and conclude that the drug does affect PSA levels. Note that the Wilcoxon signed rank test assumes that the data are quantitative; otherwise, differences between groups (eg, before minus after values) would not be meaningful.

There is another common nonparametric test for paired data, the sign test. It is similar to the Wilcoxon signed rank test except that magnitude of the raw data or the ranks do not matter, only the direction of the difference (ie, smaller, same, bigger). The test compares the number of positive and negative signed ranks and ignores pairs that are the same. The sign test is less powerful than the Wilcoxon signed rank test but is useful when the magnitude of change (or difference) is not as important as the direction of change or when the data are ordered but not interval level.

Nonparametric vs Parametric Tests

Except for the paired tests, both parametric and nonparametric tests assume random sampling from independent populations. Nonparametric tests do not assume random sampling from normal distributions as do the *t* tests. So what happens when a nonparametric test is performed with samples from normal distributions? The nonparametric tests are still valid in this case and will usually give similar results as *t* tests. However, the *t* tests can have more power (higher probability of finding a true difference) than nonparametric tests when the normality assumptions are met. However, if the data are not normally distributed or if there is doubt about the normality assumption, one should use a nonparametric test. This is especially so for small sample sizes, because it is often impossible to know what the distribution looks like from small samples.

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

Statistical tests for more than 2 samples

Jerry W. McLarty, PhD,* and Runhua Shi, MD, PhD*

INTRODUCTION

In a previous article statistical methods for comparing 2 groups were discussed. However, what if there are more than 2 groups? For example, how do you test for differences in the effectiveness of 3 treatments, say, drug A, drug B, and drug C? The 2-group *t* test and its equivalent nonparametric tests are tailored for comparing only 2 means or medians. A simple alternative comes to mind: compare drug A to drug B, drug A to drug C, and drug B to drug C. So, instead of 2 comparisons there are now 3. If there were 4 groups to test, it would require 6 two-group comparisons. For 5 groups, 10 comparisons would be required. Multiple *t* tests are not the preferred method of addressing the multiple-group comparison problem. The statistical problem with multiple *t* tests in this manner is discussed in more detail herein, and more powerful methods are introduced. Also, extension of χ^2 and other nonparametric tests to compare multiple groups simultaneously is described.

MULTIPLE COMPARISONS PROBLEM

The fundamental problem with using multiple *t* tests to compare more than 2 groups simultaneously is that the type I error rate (ie, probability of erroneously rejecting the null hypothesis by chance) is increased with each comparison. This is called the *multiple comparisons problem*. The problem can occur in a number of statistical testing situations. To illustrate how this happens, consider a study comparing the effectiveness of 3 different drugs. The null hypothesis is that the 3 drugs have an equal effect: H_0: $\mu_A = \mu_B = \mu_C$, where the μ's are the respective means for each drug. If the probability of type I error $\alpha = .05$ for a single *t* test (H_0: $\mu_A = \mu_B$ or H_0: $\mu_B = \mu_C$ or H_0: $\mu_A = \mu_C$), the actual type I error rate for all 3 *t* tests combined is approximately 0.14. This means that 14 times of a hundred we would expect to reject the null hypothesis by chance, not 5 as for $\alpha = .05$. The overall type I error rate continues to increase with more tests: for 4 *t* tests it would be approximately 0.19. It can be seen that the type I error rate can become unacceptably large, and we would have less faith in our conclusion to reject the null hypothesis and claim that one of the drugs is best.

One simple method to get around the multiple comparisons problem is to adjust the *P* value: if there are n tests, reject the null hypothesis only if $P/n < .05$. In our 3-group example, a *P* value of $.05/3 \cong .017$ or less would be called significant. This is called the Bonferroni correction, and it is commonly used. Gene array testing, for example, often performs tens of thousands of tests looking for polymorphisms linked to specific diseases. The Bonferroni corrected *P* values can be quite small in such testing (eg, 10^{-9} or lower).

The Bonferroni correction is simple to apply and is widely used, but it has been criticized as being too conservative, that is, more often than necessary it fails to reject the null hypothesis. Numerous other methods of adjusting for multiple comparisons have been proposed.

COMPARISON OF MEANS: ANALYSIS OF VARIANCE

The *t* test is a comparison of means test. As discussed in an earlier article, the *t* test can be used to compare one mean value (from a random sample) with a single number or to compare the mean from one group with the mean of another group or to compare 2 closely paired groups. The basic premise is that the difference in 2 means divided by the SE of the difference follows a well-known distribution, the *t* distribution. The *t* statistic can be seen as a signal to noise ratio (ie, the test statistic):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SE(\bar{X}_1)^2 + SE(\bar{X}_2)^2}} = \frac{\text{signal}}{\text{noise}}$$

which consists of the signal, the difference of the 2 means, and the noise, all that stuff in the denominator. This is a useful way of thinking about statistical tests: as the signal gets bigger (more difference between means), the test statistic gets larger and the *P* value is more significant. Conversely, if the noise in the system (ie, the variance or standard deviation (SD) of the samples) is large, the value of the test statistic decreases and the *P* value is less significant.

This same signal to noise analogy is the basis of a powerful statistical test, the analysis of variance (ANOVA). ANOVA in its simplest form is an extension of the *t* test to more than 2 groups. For ANOVA, the test statistic is called F and has its own probability distribution. Like the *t* test statistic, F is the ratio of 2 quantities:

F = mean square between groups/mean square within groups.

The derivation of this formula is given herein but may be skipped by readers more interested in an overview of ANOVA than computational details.

DERIVATION OF THE F TEST FORMULA

Mean Square Between Groups (Signal)

For the *t* test the difference in the 2 group means is the numerator of the formula; in ANOVA we have several group

Affiliations: *Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication February 16, 2009; Received in revised form April 2, 2009; Accepted for publication April 3, 2009.

means to compare, so the trick is to calculate the difference between each group mean and the overall mean (the mean of all the values in all groups). If the overall mean is $\bar{X}_{ALL}$, and $\bar{X}_A$, $\bar{X}_B$, and $\bar{X}_C$ are 3 group means we wish to compare, then $SS_{\text{Between Groups}} = n[(\bar{X}_A - \bar{X}_{ALL})^2 + (\bar{X}_B - \bar{X}_{ALL})^2 + (\bar{X}_C - \bar{X}_{ALL})^2]$, where n is the number of subjects in each group and SS is the abbreviation for sum of squares. If we added just the 3 differences without first squaring them, they would cancel out and sum to 0, which would give us no information about the signal. To avoid this, the differences are squared, hence the term *sum of squares*.

Only one new concept needs to be introduced, that is, mean square. If we wanted to get an average squared difference it would be intuitive to divide the SS described previously by 3 (3 group differences were obtained). However, as with SD, to get an unbiased estimate using our sampled groups, we divide the sum of squares by the degrees of freedom. One degree of freedom was used in calculating the overall mean. So, to get our average, we divide the sum of squares by the number of groups minus 1, that is, k − 1; in this example, $df = 3 - 1 = 2$. In general, the between-groups degrees of freedom ($df_{\text{Between Groups}}$) = k − 1, where k is the number of groups. Finally, we calculate the average, called the mean square (abbreviated as MS):

$$MS_{\text{Between Groups}} = SS_{\text{Between Groups}}/df_{\text{Between Groups}},$$

this is the numerator (signal) portion of our F statistic.

Mean Square Within Groups (Noise)

The goal of ANOVA is to compare means between groups. There is variability between groups and variability within each group. It is this variability within groups that leads to the denominator (noise) portion of our F statistics.

For our example with 3 groups, let's assume that the sample size is 10 individuals for each group or 30 people total. If we compute the squared difference between every data point and its group mean and then add them, we would have a sum of squares within groups:

$$SS_{\text{Within Groups}} = (X_{1A} - \bar{X}_A)^2 + (X_{2A} - \bar{X}_A)^2 + \ldots + (X_{10A} - \bar{X}_A)^2 + (X_{1B} - \bar{X}_B)^2 + (\bar{X}_{2B} - \bar{X}_B)^2 + \ldots (X_{1C} - \bar{X}_C)^2 + \ldots(\bar{X}_{2C} - \bar{X}_C)^2 + \Lambda + (\bar{X}_{10C} - \bar{X}_C)^2$$

where X_{1A}, for example, means the value of the first data point in group A, and … denotes terms omitted for brevity.

For our example, there was a total sample size of 30. Three degrees of freedom were used to calculate the 3 group means. So the overall within-group degrees of freedom is as follows: $df_{\text{Within Groups}}$ = N − k (where N is the total sample size for all groups combined and k is the number of groups). For our example, $df_{\text{Within Groups}} = 30 - 3 = 27$. In a manner similar to between groups, the within-groups mean square is as follows:

$$MS_{\text{Within Groups}} = SS_{\text{Within Groups}}/df_{\text{Within Groups}}$$

Calculation of F

Now that both our numerator (signal) and denominator (noise) have been defined, we can calculate the test statistic as follows:

$$F = MS_{\text{Between Groups}}/MS_{\text{Within Groups}}$$

With the test statistic F, the *P* value can be calculated with a computer or looked up in a table with numerator degrees of freedom k − 1 and denominator degrees of freedom N − k.

Example (Hypothetical)

Assume that diastolic blood pressure is measured in 5 subjects in each of 3 age groups: 30 to 39, 40 to 49, and 50 to 59 years. The question is, "Does diastolic blood pressure differ in these age groups?" The null hypothesis is the group means are equal. The data table looks like this:

Subject No.	Age 30–39 years	Age 40–49 years	Age 50–59 years
1	70	80	90
2	75	70	85
3	60	75	80
4	70	85	95
5	80	80	80
Group mean	71.0	78.0	86.0

The overall mean is 78.3 mm Hg.

The sum of squares between groups is as follows:

$SS_{\text{Between Groups}} = 5[(71–78.3)^2 + (78–78.3)^3 + (86–78.3)^2] = 563.3$

$MS_{\text{Between Groups}} = SS_{\text{Between Groups}}/df_{\text{Between Groups}} = 563.3/2 = 281.7$

$SS_{\text{Within Groups}} = (70–71)^2 + (75–71)^2 + \ldots + (80–78)^2 + (70–78)^2 + \ldots(90–86)^2 + \ldots (80–86)^2] = 520.0$

$MS_{\text{Within Groups}} = SS_{\text{Within Groups}}/df_{\text{Within Groups}} = 520/12 = 43.3$

Finally,

$F = MS_{\text{Between Groups}}/MS_{\text{Within Groups}} = 281.7/43.3 = 6.5$

The *P* value for F = 6.5 with 2 and 12 *df* = 0.012.

Because *P* is less than .05, we reject the hypothesis that the group means of diastolic blood pressure are equal.

Which Groups Are Different?

In this example, the F test tells us that the groups means are not all the same (ie, at least one of them is different from another). Unfortunately, it does not tell us which one(s) is (are) different. In the example, it is easy to see that the youngest age group has a much lower mean diastolic blood pressure than does the oldest, but it is not clear whether the middle age group significantly differs from the other two. To address this issue, there are several post hoc tests (tests to be used after the null hypothesis has been rejected) that can help sort this out. Post hoc tests are beyond the scope of this

discussion, but most statistical software has several of them included with the ANOVA procedure.

Comparison of ANOVA and the t *Test*

The underlying assumptions of the *t* test apply to ANOVA: the groups are assumed to be independent random samples from normal populations. In its simplest form, ANOVA is just an extension of the *t* test to more than 2 groups. If ANOVA is performed with just 2 groups, the *P* value will be identical to the *P* value of a *t* test on the same data.

Extensions of Simple ANOVA

ANOVA is a powerful statistical tool with capabilities far beyond expanding *t* test methods to more than 2 groups. It is possible to have more than one factor (in the example given age group is a factor) in ANOVA. For example, a similar study might include sex and race and even include interaction effects among them; it might be that drugs work differently for older African American men than for white men and women and African American women. In this case there would be a significant age-race-sex interaction effect. Factors in ANOVA are variables of interest that can be categorized into a few values, such as age group, sex, race, and study center. A single ANOVA test can test all of the factors at once. The basic principals of comparing between and within mean squares are the same as for the simple 1-factor ANOVA.

Analysis of covariance (ANCOVA) extends this concept to continuous variables (called *covariates*), such as age, height, and cholesterol level. Again, ANCOVA tests all factors and covariates and interactions in a single run. Generally, covariates are tested first, and if significant, the factors are tested after adjusting for the covariates. For example, if age is a covariate and sex is a factor, the effect of sex may be computed after adjusting for age, simulating the effect that all subjects are the same age.

Another powerful extension of ANOVA is repeated-measures ANOVA. In many studies, the same outcome is measured multiple times on the same individuals. For example, the effect of an allergy treatment might require measuring forced expiratory volume in 1 second (FEV_1) at weekly intervals for 5 weeks. The investigators could just compare the effect of the drug, compared with a placebo, by using the FEV_1 at week 5. However, this ignores information that may be useful from the first 4 weeks of data. Repeated-measures ANOVA is a means of using all the data from all of the time points.

It is possible that the sample sizes are not always equal for each group. However, the same methods described herein still apply, but the computation of degrees of freedom within groups is different. Most common software packages take care of this automatically. In general, ANOVA calculations are sufficiently tedious so it is not practical to do them by hand when there are so many software packages readily available. Statistical software will be discussed in a later article. Readers interested in more advanced applications of ANOVA should consult statistical textbooks.

COMPARING PROPORTIONS IN MORE THAN 2 GROUPS

Two-sample comparisons of proportions have been described in a previous article. Expansion of the χ^2 test to more than 2 samples (groups) involves no new concepts. The principle of comparing observed to expected values in each cell of a table is identical regardless of the number of samples, and the rules of thumb concerning the magnitude of expected values are the same.

Recall that for each cell in a table the comparison of the observed data to the expected data (based on the assumptions that rows and columns are independent) follows this format:

$$(\text{Observed} - \text{Expected})^2/\text{Expected}$$

The differences are squared to eliminate the sign of the difference. Otherwise the sum of differences used in the following equation would always sum to zero (ie, pluses and minuses would cancel each other out). The χ^2 test statistic is the sum of all such squared terms in the table. In mathematical symbols this becomes the following:

$$\chi^2 = \sum\sum \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} = \frac{\sum_i \sum_j (O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where i = 1,2,3, …, N_{rows} indexes rows and j = 1,2,3,…N_{cols} indexes columns.

If appropriate conditions are met, then this sum approximates a χ^2 distribution and probabilities (*P* values) can be calculated. The degrees of freedom for calculating *P* values is $(N_{rows} - 1) \times (N_{cols} - 1)$. There is no restriction on how large the number of rows and columns can be, except there must be at least 2 cells in the table. As before, no expected values can be less than 1.0, and 20% or more of the expected values must be 5.0 or greater. If these rules are violated, the approximation to the χ^2 distribution is suspect and the conclusions reached may be in error. This can often be fixed by collapsing rows or columns (ie, combining 2 or more groups into 1 group). If it is not possible to do this, there is a test called the Fisher exact test that is not restricted by small or even zero expected values. However, this can be computationally intensive, depending on the size of the table and the number of subjects. To illustrate the χ^2 test, consider the following example, an extension of the example used in an earlier section.

Assume we want to compare the gene frequencies of a gene mutation possibly involved in an allergic disease for 3 different racial groups: Asian, non-Hispanic white, and Hispanic white. The data are as follows:

Subjects	AA	AG	GG	Total No. of subjects
Asian	15 (12.2)	60 (47.1)	30 (45.7)	105
Non-Hispanic white	10 (13.8)	44 (53.3)	65 (51.8)	119
Hispanic white	3 (2.0)	4 (7.6)	10 (7.4)	17
Total	28	108	105	241

The numbers in parentheses are expected values. For tables like this, the expected values for a given cell (say the AA genotype in Asians) is column total (28) times row total (105) divided by the overall total (241). This assumes that the column values are distributed proportionately in each row.

The χ^2 for this table is 18.8, with 4 *df*. All of the expected values are greater than 1.0, and only one cell had an expected value below 5.0. The χ^2 assumptions are valid for this table. The *P* value is .001. We conclude that there are racial differences in the distribution of genotypes, with Asians having a predominance of the AG genotype.

NONPARAMETRIC COMPARISONS OF MORE THAN 2 GROUPS

As we have seen, ANOVA assumes that the dependent (outcome) variable is drawn from a normal distribution. When this assumption is not met, nonparametric methods based on ranks have been developed. The 2-sample case has been described in a previous article. The 2-sample case is called the Mann-Whitney test. An equivalent test for more than 2 groups is the Kruskal-Wallis test. The principal idea behind the Kruskal-Wallis test is to rank all the data without regard to group and then calculate an overall average rank and an average rank within each group. Then the difference between group mean ranks and the overall mean rank is calculated and summed over all groups, much like the sum of squares for ANOVA. The formula for the test statistic K is as follows:

$$K = \frac{12}{N(N+1)} \sum_{1}^{g} n_i(\bar{r}_i - \bar{r})^2$$

where N is the total number of subjects, g is the number of groups, n_i is the number of subjects in group i, $\bar{r}_i$ is the mean rank for group i, and $\bar{r}$ is the overall mean rank.

For group sizes of 5 or more, K has an approximate χ^2 distribution with $g - 1$ *df*. To illustrate, let's apply the Kruskal-Wallis test to the following example. Assume we want to compare the prostate specific antigen (PSA) levels in 3 groups of men ages 40 to 50, 50 to 60, and 70 to 80 years. This is a hypothetical example, but PSA is a good illustration: it is not normally distributed but is highly skewed.

Raw Data

Subject No.	Age 40–49 years	Age 50–59 years	Age 60–69 years
1	0	0.4	1.2
2	0.1	0.6	2.0
3	0.2	0.8	3.0
4	0.3	1.0	2.4
5	0.5	1.4	4.6

Ranked Data

Subject No.	Age 40–49 years	Age 50–59 years	Age 60–69 years
1	1	5	10
2	2	7	12
3	3	8	14
4	4	9	13
5	6	11	15

The overall mean rank is 8.0. There are 3 groups and a total of 15 subjects. So,

$$K = \frac{12}{15(16)}[5(3.2 - 8)^2 + 5(8 - 8)^2 + 5(12.8 - 8)^2]$$

K =11.52, $df = 3 - 1 = 2$, $P = .003$.

So we reject the null hypothesis of no age difference in PSA.

The problem of which groups are different remains. Unfortunately, the methods are more limited than for ANOVA. A simple alternative, after determining from the Kruskal-Wallis test that the groups are not the same, is comparing pairs of groups using the Mann-Whitney test and reducing the type I error rate from 0.05 to 0.05 divided by the number of comparisons (Bonferroni correction). In this example, we would reject the pairwise null hypothesis with $P < .05/3$ or .0167.

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

Correlation and regression analysis

Runhua Shi, MD, PhD,* and Steven A. Conrad, MD, PhD†

INTRODUCTION

The preceding 2 articles have focused on the comparison of 2 or more samples for the purpose of testing for a difference among the samples with respect to 1 or more outcome variables. In contrast to *outcome*, it may be desirable to determine the *relationship* between 2 or more variables. The methods introduced in this article include correlation and regression.

Correlation analysis assesses the linear relationship between 2 variables, providing a measure of both the strength and direction of the relationship. Correlation makes no assumption on causality in the relationship. It assumes only a linear relationship, and variables with a strong nonlinear relationship may show poor or absent correlation. To help identify the type of relationship between variables, visual inspection of a scatterplot is invaluable. Correlation can be performed on both parametric and nonparametric variables. The most commonly used parametric method is the Pearson product-moment correlation. Two nonparametric methods are in common use, including the Spearman rank order correlation and Kendall τ methods. *Partial correlation* provides for a measure of correlation after controlling for the effects of variables other than the 2 primary variables. In certain situations, the correlation relationship can be linear to a certain extent beyond which it may disappear or remain linear but at a different degree.

Regression analysis assesses the relationship between 1 dependent (observed) variable and 1 or more independent (explanatory) variables, with an implied causal relationship. Regression goes beyond correlation by inferring relationships between variables, allowing modeling of causal relationships, and predicting the value of the dependent variable from a given value of independent variables(s). Unlike correlation analysis, which makes few assumptions, regression analysis is based on a number of underlying assumptions. Regression analysis includes both linear and nonlinear regression. *Linear regression* involves a linear model, which is linear with respect to its parameters. Linear regression models may be simple (a single independent variable) or multiple (2 or more independent variables). *Nonlinear regression* deals with exponential, power, or more complex relationships.

Logistic regression extends regression analysis to include dependent variables, which may be dichotomous (binary) or discrete (multinomial) instead of continuous, and independent variables, which may be a combination of continuous, discrete, and/or dichotomous. The underlying assumptions are considerably relaxed from those of linear regression. As in linear regression, logistic regression models may be simple or multiple.

CORRELATION ANALYSIS

The correlation coefficient (r) is a measure of the strength of the linear relationship between 2 variables. A positive correlation indicates that as one variable increases the other increases also. A negative correlation indicates that one variable increases as the other decreases. A value of 1 indicates a perfect correlation (ie, the first variable is an exact linear function of the second with a positive relationship). A value of -1 indicates an exact negative linear relationship. Values between 0 and these extremes indicate increasing strength of association in a positive and negative relationship, respectively. It is generally accepted that the correlation is considered weak if $r \leq 0.4$, moderate if $0.4 < r < 0.8$, and strong if $r \geq 0.8$.

For illustration purposes, Table 1 shows a data set of annual income, unemployment rate, and weekly food expenditures according to years of education. In Figure 1, we can see a positive correlation between education level and annual income, a negative correlation between education level and unemployment rate, and an absent correlation between education level and weekly expenditures for food. The correlation coefficients and significance tests for these examples are discussed herein.

PEARSON PRODUCT-MOMENT CORRELATION

The Pearson product-moment correlation coefficient (Pearson r) is a parametric measure of linear association between 2 variables. Its underlying assumptions include the variables each being normally distributed and their joint distribution being *bivariate normal*. The bivariate normal distribution is

Table 1. Hypothetical Data on Education, Annual Salary, Unemployment, and Weekly Food Expenses

Education level, y	Annual salary (in thousands), $	Unemployment rate, %	Weekly food expenditure, $
8–9	13.0	11	342
10–11	19.7	7.1	347
12–13	26.0	4.0	339
14–15	31.7	2.5	341
16–17	40.1	1.9	314
18–19	50.0	1.6	352
20–21	62.4	1.4	346
≥22	71.0	1.3	348

Affiliations: *Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana; †Departments of Medicine, Emergency Medicine, and Pediatrics, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication January 28, 2009; Received in revised form April 9, 2009; Accepted for publication April 12, 2009.

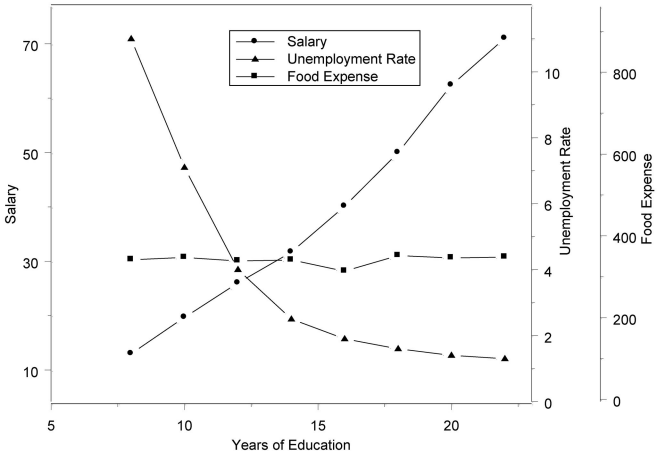


Figure 1. Correlation between annual salary, unemployment rate, and weekly food expenditure according to the years of education (data in Table 1).

an extension of the normal distribution to 2 variables. The bivariate probability density function is a 2-dimensional surface with a center peak and gaussian-type decrease as the distance from the center increases. In most practical applications, if the 2 variables satisfy the normal distribution assumption, they usually satisfy the bivariate normal assumption.

The Pearson product-moment correlation for a sample estimates the correlation for the population. The formula for the sample Pearson product-moment correlation is as follows:

$$r_{xy} = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where $\bar{x}$ is the sample mean of variable x and $\bar{y}$ is the sample mean of variable y. When the data represent the population instead of just a sample, the correlation coefficient is designated as ρ_{xy} and is calculated from a similar formula:

$$\rho_{xy} = \frac{\sum_i ((x_i - \mu_x)(y_i - \mu_y))}{\sqrt{\sum_i (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$

where μ_x and μ_y represent the population means for the x and y variables. Typically, the population means are not known and the sample correlation is what is calculated.

The *coefficient of determination*, calculated as r^2, is a derived statistic that provides an indication of the strength of the relationship between the 2 variables. The value of r^2 corresponds to the amount of variation in one of the variables explained by the other. For example, a value of r of 0.8 indicates that 64% (0.8^2) of the variation in the first variable is accounted for by knowing the second (or vice versa).

When interpreting the value of the correlation coefficient, it should be recognized that the relative importance of the value depends on the sample size. In small sample sizes, a magnitude of 0.1 would generally not be considered substantial, but in very large samples, this value might well represent a significant correlation. This dependence is evident when formally testing for significance. Like other statistical measures, a correlation coefficient can also be tested for statistical significance. The value of the sample correlation coefficient r is assumed to deviate around the population coefficient ρ with a given distribution. This distribution is skewed if r is tested against values toward -1 or 1. However, when testing that r is different from zero (the most common scenario), this distribution is symmetrical and can be estimated using the t distribution, where

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

with $(n-2)$ *df*. As noted herein and evident in the equation, the sample size (n) has a substantial impact on the value of the t statistic independent of $|r|$.

As an example, consider the correlation between body weights in spouses. We assume that weight is normally distributed in both groups. Suppose 50 spouse pairs were weighed, with a resulting correlation coefficient of 0.6. We can test the null hypothesis H0: $\rho = 0$ vs the alternate hypothesis H_1: $\rho \neq 0$. This value of r and n give a t value of 5.20, which is statistically significant, $P = .02$, so we can conclude there is a positive moderate correlation between the weight of spouses in this sample.

If there is a need to test the hypothesis that ρ is equal to a specific value other than zero, the t statistic described cannot be used. In this case, the Fisher z transformation can be helpful. This is a more general approach that can also be used to test the hypothesis that ρ is equal to zero. The initial step is to transform the correlation to eliminate the effect of skewness:

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

This transformed variable to natural logarithm (ln) approximates a normal distribution, so that a z statistic can be calculated to test the null hypothesis that $z_r = z_0$:

$$z = \frac{z_r - z_0}{SE}$$

$$= \frac{z_r - z_0}{\sqrt{\frac{1}{n-3}}}$$

The value of z can be evaluated with the cumulative normal distribution function to obtain the level of significance, either 1-tailed or 2-tailed.

An alternative approach is to calculate a confidence interval (CI) around the transformed z value:

$$CI = z_r \pm z_{critical} \times SE$$

which for the 95% CI would be

$$CL = z_r \pm 1.96 \times \sqrt{\frac{1}{n-3}}$$

Using the sample data set in Table 1 and graphed in Figure 1, the following Pearson coefficients, CI, and *P* value can be calculated:

	Pearson coefficients	95% CI	*P* value
Education level vs salary	0.992	0.947 to 0.998	<.0001
Education level vs unemployment	−0.872	−0.973 to −0.384	.0027
Education level vs food expenses	0.121	−0.643 to 0.757	.785

The first reveals a very strong positive correlation, the second a strong negative correlation, and the third a weak nonsignificant correlation.

SPEARMAN RANK-ORDER CORRELATION

The nonparametric measure of Spearman rank-order correlation (Spearman ρ; in contrast to the parametric correlation r) is based on the ranks of the data values rather than the raw data values. Unlike the Pearson correlation, it does not require the underlying assumption of normality and should be the correlation used in this circumstance. It is particularly useful when the data represent ordinal or qualitative variable or contain outliers.

The Spearman correlation is a special case of the Pearson correlation in which the values x and y are replaced with their rankings. This leads to a formula based on the Pearson coefficient:

$$r_s = \frac{\sum_i ((R_x - \bar{R}_x)(R_y - \bar{R}_y))}{\sqrt{\sum_i (R_x - \bar{R}_x)^2 \sum (R_y - \bar{R}_y)^2}}$$

where R_x and R_y are the ranks of the x and y variables, respectively. A more efficient formula is as follows:

$$\mathbf{r_s} = 1 - \frac{6 \sum \mathbf{d_i^2}}{\mathbf{n}(\mathbf{n}^2 - 1)}$$

where d_i^2 is the square of the difference between the corresponding ranks of x_i and y_i, and n is the number observations. In the case of tied ranks, this formula cannot be used, and the correlation must be calculated by an alternate method, such as the Pearson correlation on the ranks, as introduced in the previous equation.

The method to test whether the Spearman correlation is significant is similar to method for Pearson correlation. In Spearman rank-order correlation, 1 or both of the variables may be ordinal or have a distribution that is far from normal. Significance tests based on the Pearson correlation coefficient will then no longer be valid, and nonparametric analogs to these tests are needed. The Fisher *z* transformation introduced for the Pearson correlation can be used to perform hypothesis testing or derive confidence limits.

OTHER CORRELATION COEFFICIENTS

Other correlation coefficients that are worth mentioning are the Kendall τ correlation, Cronbach coefficient α, and partial correlation.

Kendall τ is a nonparametric measure of association based on the number of concordances and discordances in ranked, paired observations. Three versions, τ-a, τ-b, and τ-c, differ in the way ties are handled, with τ-b being the most common. This correlation method is used in the field of testing new methods against old methods regarding the sensitivity and specificity. Values range from −1 to 1, with 0 reflecting absence of correlation.

Another type of correlation coefficient is the Cronbach coefficient α. This method is commonly used in reliability studies to determine the internal consistency of a test or the average correlation of items within the test. Interrelated items may be summed to obtain an overall score for each participant. The larger the overall α coefficient, the more likely that items contribute to a reliable scale. A value of ≥0.70 is an acceptable reliability coefficient; smaller reliability coefficients are seen as inadequate. However, this varies by discipline.

When more than 2 related variables are measured, one could measure the *partial correlation* of 2 variables while controlling for other variables. A partial correlation measures the strength of a relationship between 2 variables, while controlling the effect of other variables. For example, age, weight, and height were measured for a group of middle school students; a partial correlation can be measured for weight and height while controlling for age.

REGRESSION ANALYSIS

Regression analysis refers to a set of methods for modeling of numerical data. A regression model consists of a defined relationship between a dependent (observed or response) variable and 1 or more independent (explanatory) variables. An assumption is that there is a causal or controlling relationship between the dependent and independent variables. Statistical models can be classified in several ways, such as the number of independent variables (simple vs multiple regression), the linearity of the parameters (linear vs nonlinear regression), or the underlying distribution of the variables (ordinary vs logistic regression). This section will introduce each of these types of regression.

The form of a *regression model* can be depicted as the following generic equation:

$$Y = f(\mathbf{X}, \boldsymbol{\beta})$$

which simply states that the single dependent variable Y depends on a set of 1 or more independent variables $\mathbf{X} = X_1 \ldots X_n$ and a set of parameters to be fit: $\boldsymbol{\beta} = \beta_0 \ldots \beta_n$. Where f represents a model function (eg, linear, polynomial). The corresponding *regression equation* includes an error term ε_i that represents the deviation of observations from the regression model:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}, \varepsilon_i)$$

Specific forms of these generic equations for the various types of regression will be introduced in the sections that follow.

Regression analysis is often applied (inappropriately) without respecting the underlying assumptions. To use regression, the following assumptions must be considered:

- The sample subjected to regression analysis should be representative of the population it is drawn from if the purpose of the regression is to develop prediction equations. Many data sample collections are convenience samples, so their ability to represent the population must be critically evaluated.
- If more than one independent variable is used, the variables should be linearly independent. A variable is dependent on another if a change in one produces an expected change in the other. An example would be the QRS vector magnitude in leads I, II, and aVF of the electrocardiogram. Any 2 of these precisely determines the third; therefore, the third measurement is linearly dependent on the other 2, and its use in a regression model would unnecessarily complicate the regression or could even invalidate the regression.
- The independent variables must not have error associated with their measurement, or the error should be insignificant compared with that in the dependent variable. This is perhaps the most common assumption violated in practice. The ideal situation is when the independent variable is experimentally (and precisely) set, rather than randomly observed, followed by measurement of the dependent variable. If significant error is associated with both dependent and independent variables, then alternative regression techniques such as orthogonal regression should be used.
- The variance of the error should be consistent across the range of independent variables. If a scatterplot of the data shows the spread in the dependent variable increasing with increasing values of the independent variable, this assumption is violated. In this case, one should consider whether a transformation of variables or the use of a weighted least squares technique (not discussed in this article) is appropriate.
- The error associated with measurement should be a random variable (randomly distributed around zero). This assumption is usually not assessable a priori but can be examined. This variable should follow a multivariate normal distribution, which is difficult to determine, so the presence of a multivariate normal distribution is usually assumed if the independent variables are each assumed to be normally distributed.

METHOD OF LEAST SQUARES

The regression techniques given herein use the *method of least squares*. This method provides values of the regression parameters by minimizing the sum of the squared deviations of the observations in the Y direction to produce the *prediction* model:

$$\hat{Y}_i = f(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ is the set of best-fit coefficients, $\hat{Y}_i$ is the predicted value for the given set of independent variables $\mathbf{X}_i$, and f is the model function (eg, linear, polynomial). In other words, the $\hat{\boldsymbol{\beta}}$ parameters are chosen so that the regression equation best fits the data, where "best" means that the sum of the distances between each observed data point and the regression line, squared, is the smallest possible sum. The best-fit coefficients are calculated from equations appropriate to each type of regression analysis. Linear least squares can be performed using linear algebraic techniques, in which the estimation equation is as follows:

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{y}$$

For the reader unfamiliar with matrix notation, X^{T} is the transpose of the X matrix and $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$ is the inverse of the matrix $(\mathrm{X}^{\mathrm{T}}\mathrm{X})$. Here, the matrix $\mathbf{X}$ contains the x values for each parameter, with each row representing a single observation, and the column vector $\mathbf{y}$ indicating the corresponding y values. Common statistical packages are capable of solving these linear equations, as well as nonlinear regressions equations that require other solution approaches. A simple example of this calculation without matrix notation is given herein.

SIMPLE LINEAR REGRESSION

When only one independent variable with linear coefficients exists, we have the case of simple linear regression, with each observation indicated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

where the β_0 is the intercept, β_1 is the slope, and ε_i represents the statistical error for i^{th} data point. The least squares approach to solving for estimates of the parameters for this simple case can be simplified using Cramer equations:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

As an example, let us determine if annual income (response variable Y) can be predicted by a linear function of education in years (regressor variable X). One can estimate β_0, the intercept, and β_1, the slope, using these equations. Table 2 gives the results of simple linear regression analysis performed in SAS statistical software on education level and salary in Table 1. The overall model goodness of fit is significant ($F = 48.685$, $P < .0001$), indicating that the

Table 2. Summary Output of Simple Linear Regression Using Data in Table 1[a]

Analysis of Variance[b]					
Source	***df***	**Sum of squares**	**Mean square**	**F value**	***P* > F**
Model	1	2915.83339	2915.83339	362.48	<.0001
Error	6	48.26536	8.04423		
Corrected Total	7	2964.09875			

Parameter Estimates					
Variable	***df***	**Parameter estimate**	**SE**	***t* value**	***P* > |t|**
Intercept	1	−23.25357	3.43206	−6.78	.0005
Education	1	4.16607	0.21882	19.04	<.0001

[a] The REG procedure, model 1. Dependent variable is salary.
[b] R^2, 0.9837; and adjusted R^2, 0.9810.

model explains a significant portion of the variation in the data. The value for R^2 of 0.984 indicates that education alone accounts for 98.4% of the variation in annual salary. The parameter estimates are $\hat{\beta}_0 = -23.3$ and $\hat{\beta}_1 = 4.17$, respectively. The table also contains the t statistics and the corresponding P values for testing whether each parameter is significantly different from zero. The P values ($t = -6.78$, $P = .0005$, and $t = 19.04$, $P < .0001$) indicate that the intercept and education parameter estimates, respectively, are highly significant. Finally, the fitted model is as follows:

$$\text{Annual Salary} = 4.17 \times \text{Education} - 23.3.$$

For illustration, Figure 2 displays the linear relationship between annual income and the education level.

MULTIPLE LINEAR REGRESSION ANALYSIS

Many studies in the biomedical sciences involve more than one explanatory variable and can involve dozens or hundreds of explanatory variables. Although univariate analysis (simple linear regression) can be repeated on each of the explanatory variables, this approach does not take into account the variation due to the additional variables and reduces the ability to find any single explanatory variable as significant. Multiple linear regression controls for the contribution of each of the explanatory variables. Assuming each of the explanatory variables contributes to the model variation, multiple regression improves the chances of finding the overall model significance.

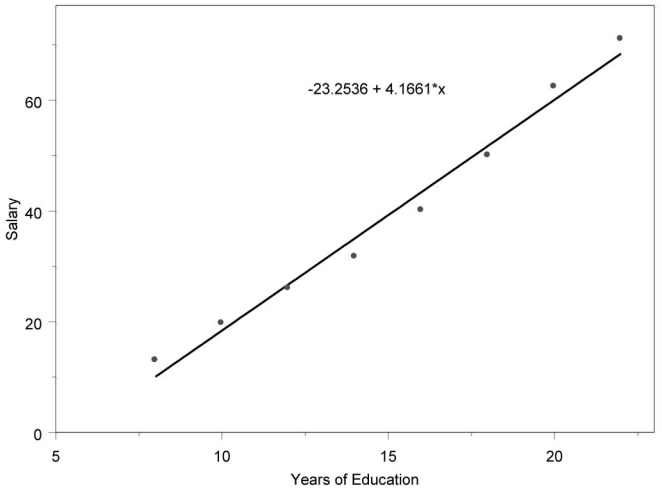


Figure 2. Linear regression results of the annual salary vs years of education.

The observations in a multiple linear regression model are indicated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in} + e_i$$

for n explanatory variables and $n + 1$ parameters and where e_i is the statistical error for i^{th} data point. The equations for analysis of multiple linear regression are more complex than the simple Cramer equations used for simple linear regressions. The calculation approach is a based on the linear algebraic approach introduced herein and is beyond the scope of this article.

The multiple linear regression model assumes that the explanatory variables are independent of each other. When there is some dependence, then an interaction results. In this case, interaction terms can be determined using one of the set of *general linear models*, which are more general than the equations presented herein and allow for the evaluation of interactions. For example, if a model were to include blood pressure and potassium intake, then a more general model should be used because there is an interaction between these 2 variables. The reader is referred to statistical textbooks for a review of general linear models.

POLYNOMIAL REGRESSION

Data with a single explanatory variable may demonstrate a curvilinear rather than a linear relationship. An example would be the relationship between education level and unemployment rate in Table 1 and Figure 1. Although this could be explained by a nonlinear equation (see the discussion on nonlinear regression), often a polynomial relationship that includes higher powers of the independent variable may provide a good fit:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_n X_i^n + e_i$$

Although at first glance this appears to be a nonlinear equation, it is linear in its *parameters* (the β terms) and thus can be approached as a variation of multiple linear regression. All that is required is to create the additional independent variables by computing the powers of the explanatory variable. Analysis then proceeds as for multiple linear regression.

An example of the use of polynomial regression for the unemployment vs education data in Table 1 is given as SAS output in Table 3. This model included the linear and quadratic terms and resulted in a better overall model goodness of fit (F = 119.6) than for the simple linear regression model (F = 19.1, output not shown), with all 3 parameters showing statistical significance. The plot of the fitted equation is given in Figure 3, where it can be seen that the polynomial fit is better than a linear fit.

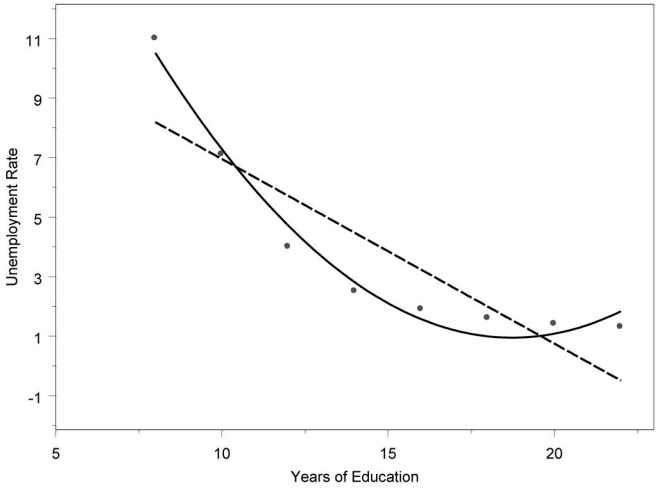


Figure 3. Polynomial regression line reflecting the analysis in Table 3 (solid line), with a linear regression line shown for comparison (dotted line).

LOGISTIC REGRESSION

Linear regression is applicable to quantitative variables. However, for a binary outcome variable, such as the prediction of developing cancer or having a stroke, traditional linear regression is often not appropriate. Often these models have binary (dichotomous), ordinal (discrete), or nominal (qualitative) explanatory variables, such as sex or stage of cancer. Logistic regression analysis is used to investigate the relationship between a binary, ordinal, or nominal response and a set of explanatory variables that may be continuous, ordinal, or binary.

The logistic function describes a sigmoidal curve that ranges from 0 to 1, as the dependent variable ranges from $-\infty$ to $+\infty$, and has a value of 0.5 when the dependent value is 0. Mathematically, it takes the following form:

$$\theta(z) = \frac{1}{1 + e^{-z}}$$

where z is the dependent variable, θ represents the probability (0 to 1), and e is a constant 2.718. Think of this function as approximating a step function with value 0 for negative (or low) values of the dependent variable and 1 for the positive (or high) values. For logistic regression with a binary outcome, we replace z with the linear equation of dependent variables:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

This yields the final form of the logistic equation, and the equivalent form used by some authors:

$$\theta(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

Conversion to a linear equation in the parameters by logarithmic transformation results in an alternative form:

Table 3. Summary Output of Polynomial Regression Using Data in Table 1[a]

Analysis of Variance[b]					
Source	***df***	**Sum of squares**	**Mean square**	**F Value**	***P* > F**
Model	2	83.16238	41.58119	119.65	
Error	5	1.73762	0.34752		
Corrected Total	7	84.90000			
Parameter Estimates					
Variable	***df***	**Parameter estimate**	**SE**	***t* value**	***P* > |t|**
Intercept	1	30.09286	2.42679	12.40	<.0001
Education	1	−3.11131	0.34413	−9.04	.0003
Education square	1	0.08304	0.01137	7.30	.0008

[a] The REG procedure, model 1. Dependent variable is unemployment.
[b] R^2, 0.9795; and adjusted R^2, 0.9713.

$$\log_e\left[\frac{\theta(x)}{1-\theta(x)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Linear logistic regression models can be fit to data by the method of maximum likelihood and are offered in statistical software packages.

NONLINEAR REGRESSION

The relationship among data variables may not be linear and may be best explained by a *nonlinear model*. Common nonlinear models include exponential and power models, but the investigator may have knowledge that a more complex relationship underlies the observed data. Fitting nonlinear models to data requires the use of iterative techniques to minimize the sum of squared residuals rather than the direct solutions available for linear models. The reader is referred to relevant statistical textbooks.

CONCLUSION

In this chapter, Pearson correlation and several nonparametric correlation methods were introduced. Pearson product-moment correlation is a parametric measure of a linear relationship between 2 variables. For nonparametric measures of association, Spearman rank-order correlation uses the ranks of the data values and Kendall τ-b uses the number of concordances and discordances in paired observations. A partial correlation provides a measure of the correlation between 2 variables after controlling the effects of other variables. Correlation coefficients can tell us if 2 variables are related, the direction of the relationship, and whether the relationship is significant. Correlation does not necessarily mean causation (but it could). Regression can tell us more about the relationship and be used to predict one from the other, and multiple regression can adjust for multiple confounding variables simultaneously. Simple and multiple linear regression methods were discussed, including polynomial regression. Logistic regression models were also presented.

REFERENCES

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

Requests for reprints should be addressed to:
Runhua Shi, MD, PhD
Department of Medicine
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-4228
E-mail: RSHI@LSUHSC.EDU

Measurements of outcome

Steven A. Conrad, MD, PhD,* and Jerry W. McLarty, PhD†

INTRODUCTION

Decisions on the interpretation of medical tests and the adoption of new therapies are part of the foundation of clinical medicine. New medical tests are constantly arising, particularly as new biomarkers arise from increased understanding of the genomics of disease. *Diagnostic tests* represent that type of medical test that is used to diagnose the presence or progression of disease. Other types of laboratory tests are used to aid in the provision of medical care, such as *therapeutic drug level monitoring*, rather than in the diagnosis of disease. The focus of this discussion is on diagnostic laboratory tests, but the principles apply to other types of medical tests as well.

An important feature of many diagnostic tests is that the test measure some component involved in the pathophysiology of, or the body's response to, the disease in question. An example of the former is the measurement of serum iron for the diagnosis of iron deficiency anemia. Because iron is required for the production of erythrocytes, it is part of the pathophysiology. An example of the latter is the measurement of white blood cell count as a marker of infection because these cells are involved in the response to infection through their role in the killing and clearance of bacteria. Other diagnostic tests rely on measurement of a marker not known to be involved in the pathophysiology but nonetheless demonstrating a correlation with the disease in question. Despite the absence of demonstrable pathogenesis, these tests are still useful if their performance is adequate. An example of this is angiotensin-converting enzyme level in sarcoidosis and other chronic diseases, diseases in which the cause is currently unknown.

DIAGNOSTIC TEST PERFORMANCE

Although not often overtly obvious, the evaluation of a diagnostic test must consider both the performance of the test with respect to the analyte (eg, measurement of C-reactive protein [CRP] in the serum) and the interpretation of that result in the context of the presence or absence of disease (eg, the use of CRP to diagnose infection). The former will usually be considered in the context of the performance measures of *accuracy* and *precision* of a particular assay, whereas the latter focuses on the evaluation of using the result of the assay in a binary classifier (eg, a disease is present or absent) with measures that include sensitivity, specificity, and others.

ACCURACY AND PRECISION

Accuracy, also known as validity, is a measure that indicates how well a test result reflects the actual concentration present in the sample. An accurate value will be close to the actual value. Precision, also known as reliability, is a measure that indicates the reproducibility of repeated measures. On repeated measurements, a precise test will have small variability.

These concepts are frequently described in terms of a marksman hitting a target (Fig 1). An accurate and precise test will have a tight grouping over the center of the bull's-eye. An accurate but imprecise test will have a wide grouping centered over the bull's-eye. A precise but inaccurate test will have a tight grouping, but the group center will miss the bull's-eye. Finally, an inaccurate and imprecise test will have a wide grouping that misses the bull's-eye.

Accuracy and precision are reported in terms of the mean and variance of the result bias (Fig 2). Bias is defined as difference between the true, or reference, value and a measurement result. Precision is inversely related to the variance and is sometimes quantitatively defined as its reciprocal, although it is usually used in a qualitative context.

Accuracy is also used to describe the performance of a test result in a binary classifier for disease diagnosis. A binary classifier maps an input variable (eg, test result) into 1 of 2 classifications (eg, presence or absence of disease). The input variable is also binary (ie, positive or negative test result). The procedure for determining a cutoff value that defines a positive test result is the subject of a later section.

A generic binary classifier is shown in Figure 3. The diagnostic test result, recorded as positive or negative, is indicated on the left. The true presence or absence of the disease or condition according to an independent gold standard is indicated on the top. The 4 resulting classifications are given in the center 4 cells. Various measures of performance are given in the bottom row and right-hand column. Accuracy in a binary classifier is defined as the proportion of all correct classifications (true-positive and true-negative results) of all observations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are perhaps the 2 most commonly used measures of binary classifier performance. *Sensitivity* is

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

Disclosures: Authors have nothing to disclose.

Received for publication February 22, 2009; Received in revised form March 19, 2009; Accepted for publication March 20, 2009.

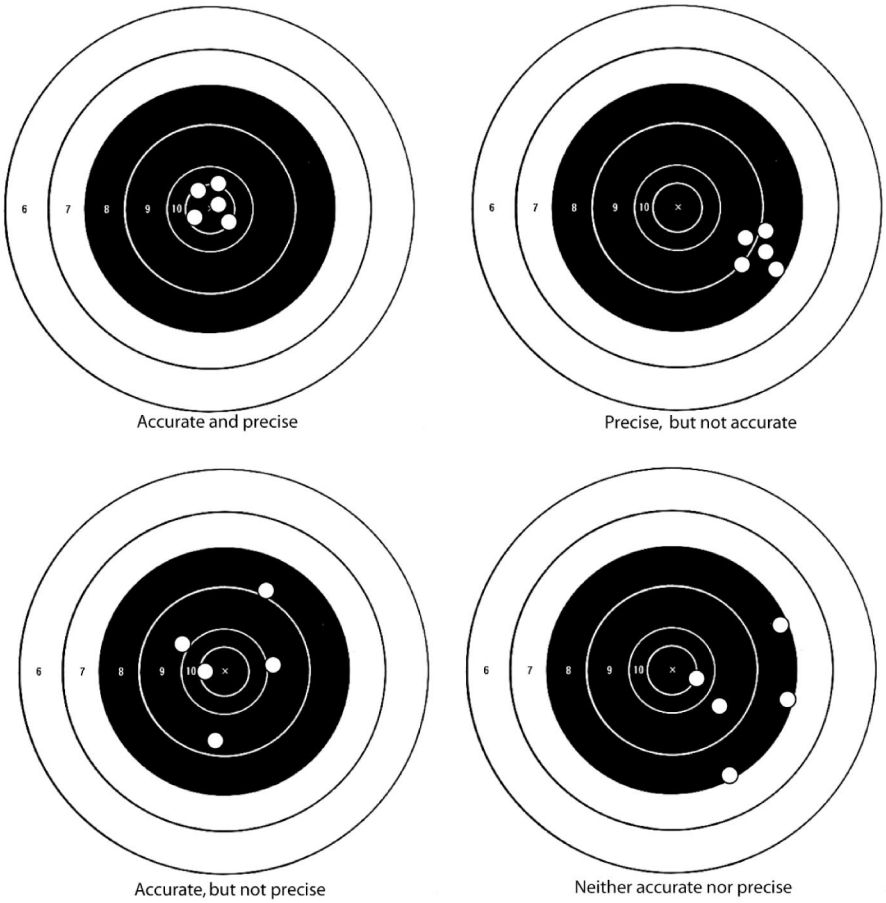


Figure 1. Schematic representation of the concepts of accuracy and precision.

a measure of the ability of a diagnostic test to detect the presence of a disease or condition when that condition is known to exist on the basis of an independent gold standard evaluation. It is calculated as the proportion of true-positive results among all patients with the disease:

$$Sensitivity = \frac{TP}{TP + FN}$$

A high sensitivity is a desirable characteristic so that the disease does not go undiagnosed. A hypothetical example is given in Figure 4. The diagnostic test is CRP, with a cutoff value of 5 mg/dL, above which the test result is considered positive. The condition to be diagnosed is bacteremia, with blood cultures used as the gold standard. With this classifier, the sensitivity is 0.72, meaning that if bacteremia is present, the CRP will be positive in 72% of cases. Sensitivity does not reflect anything about patients *without* the disease, so it is possible that a positive test result can occur in many patients without the bacteremia. Therefore, a negative result in a highly sensitive test is most useful for ruling out bacteremia.

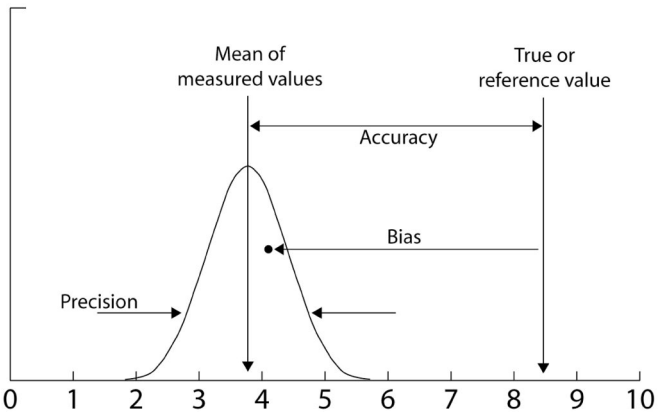


Figure 2. Schematic representation of the concepts of accuracy and precision. *Bias* is a term that indicates the difference between a *single measurement* and the true or reference value. *Accuracy* is the deviation of the *mean of a set of repeated measures* from the true or reference value. The further the mean from the reference value, the lower the accuracy. *Precision* is inversely related to the variance of the set of measured values. A higher variance represents a lower precision.

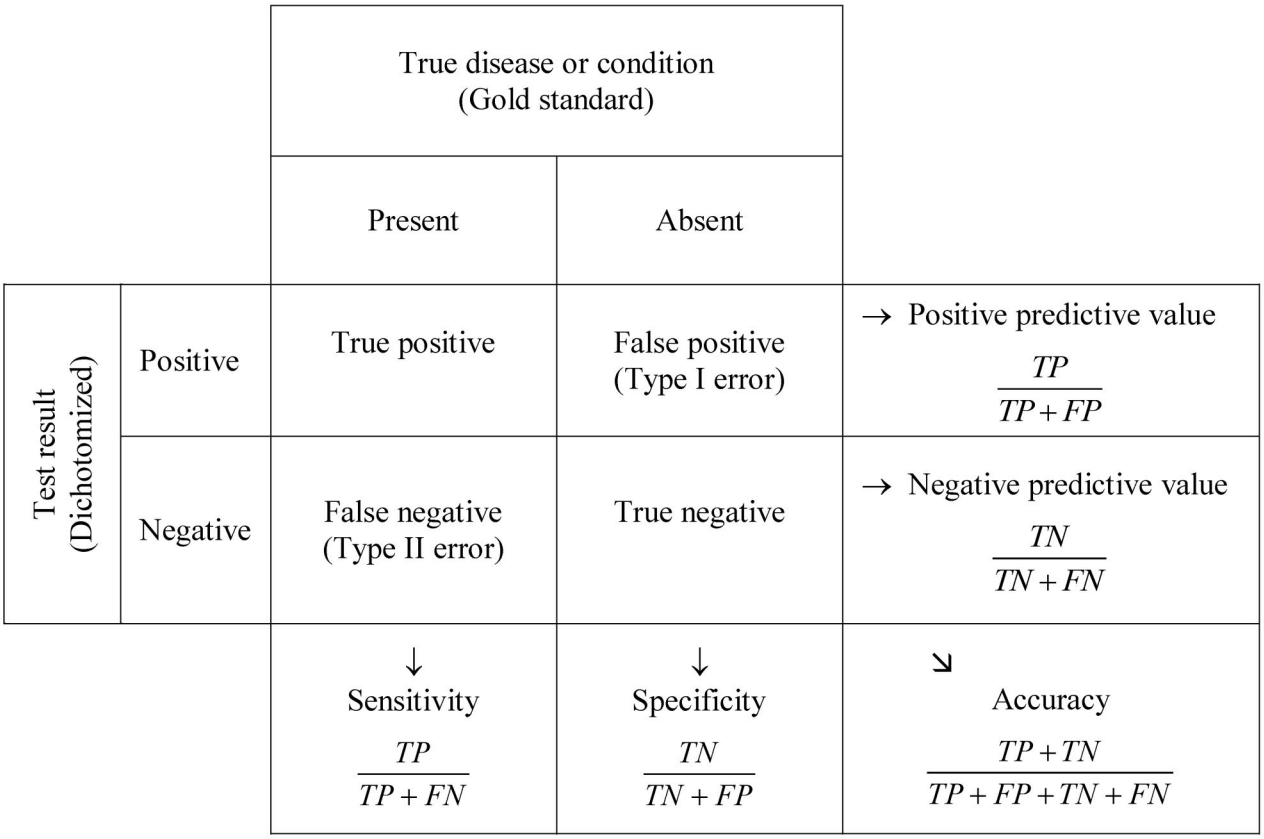


Figure 3. Generic binary classifier. The classifier columns on the left represent the diagnostic test result indicated as a positive or negative test result. The classifier rows on the top represent the presence or absence of the disease or condition as determined by a gold standard. The classifier outputs consist of 4 center states described as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The values in the bottom row and right-most column represent performance measures of the classifier, which are discussed in more detail in the text.

Specificity is a measure of the ability of a diagnostic test to exclude the presence of a disease or condition when that condition is known to be absent. It is calculated as the proportion of true-negative results among all patients without the disease:

$$Specificity = \frac{TN}{TN + FP}$$

A high specificity is desirable because we want to avoid diagnosing a condition that does not exist and perhaps instituting an unnecessary treatment. In the CRP example, the specificity is 0.80, meaning that we will wrongly diagnose bacteremia (type I error) in 20% of cases. The higher the specificity, the less likely we are to initiate unnecessary antibiotics. Specificity does not reflect anything about patients *with* the disease, so it is possible that a negative test result can occur in many patients with bacteremia. As a result, a positive result in a highly specific test is most useful for ruling in bacteremia.

Although it is desirable that a test have both a high sensitivity and high specificity, in practice this is rarely possible. In a given test the 2 tend to be inversely related, that is, a highly sensitive test tends to have a low specificity. Lowering the cutoff value that defines a positive test result will increase sensitivity and decrease sensitivity and vice versa. A topic of discussion later in this review (receiver operator characteristic [ROC]) will examine how to best balance these characteristics for a test in choosing a cutoff value.

Highly sensitive tests serve a major role as screening tests, which tend to catch most patients with the disease but also a large number without. A positive result in a sensitive screening test can then be confirmed by a different, more specific test.

POSITIVE AND NEGATIVE PREDICTIVE VALUES

The measures of sensitivity and specificity described herein reflect the performance of a test when the disease or condition is known. The health care professional, however, is presented with only the test result and wishes to know, given that result,

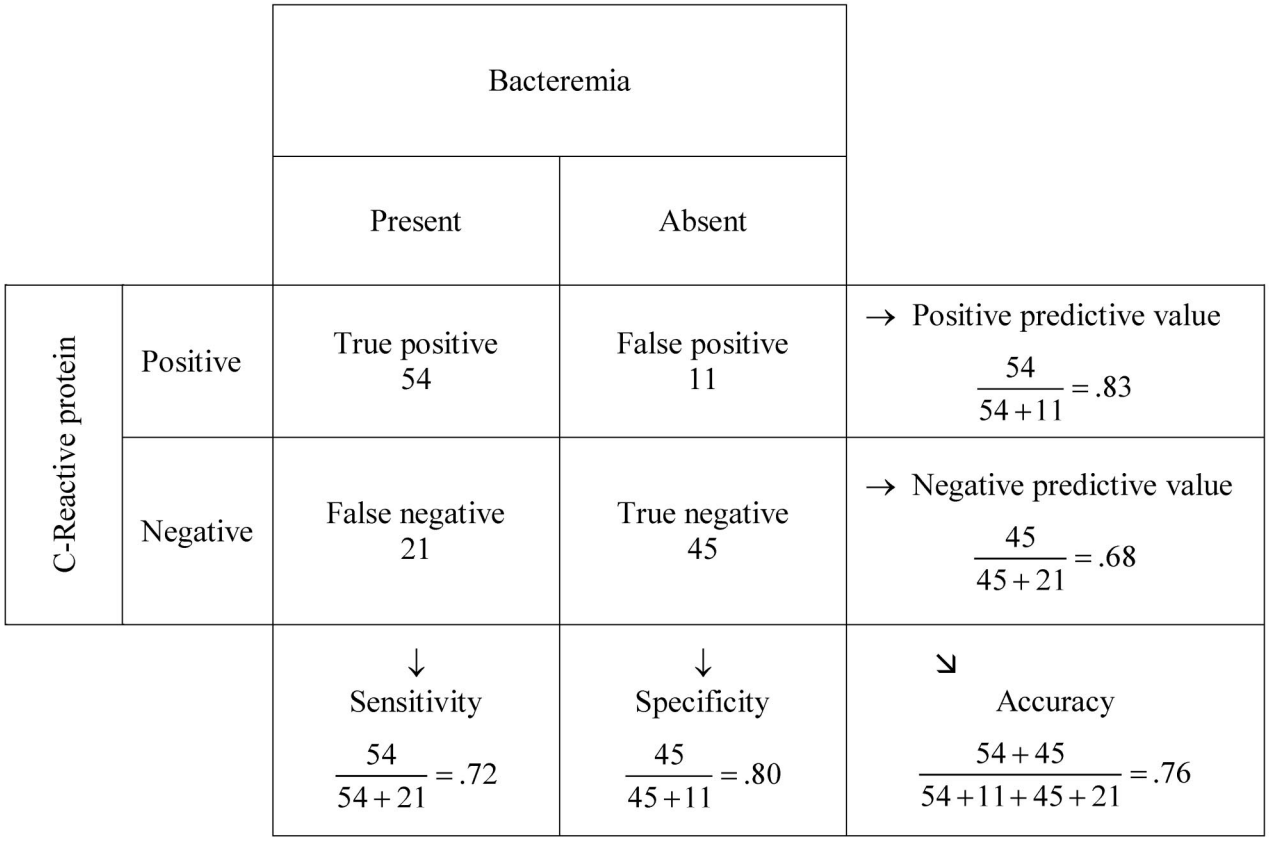


Figure 4. Example binary classifier for the use of C-reactive protein for the diagnosis of bacteremia.

how well the test predicts the disease. For this purpose the positive predictive value (PPV) and negative predictive value (NPV) provide useful information.

The PPV of a test indicates the probability that the disease is present when the test result is positive. It is calculated as the proportion of true-positive results among all positive test results (true-positive results plus false-positive results, Fig 3):

$$PPV = \frac{TP}{TP + FP}$$

A PPV of 0.83 in the example in Figure 4 indicates that a CRP above the cutoff value of 5 mg/dL correctly predicts bacteremia 83% of the time. A high PPV is desirable so that when treatment is initiated on the basis of a positive test result, the chance of providing the treatment to those without the disease is low.

Similarly, the NPV of a test indicates the probability that the disease is absent when the test result is negative. It is the proportion of true-negative results among all negative test results (true-negative results plus false-negative results, Fig 3):

$$NPV = \frac{TN}{TN + FN}$$

For the CRP result in Figure 4, an NPV indicates that a CRP below the cutoff value of 5 mg/dL correctly excludes bacteremia 68% of the time. A high NPV is desirable so that when treatment is withheld on the basis of a negative test result, the number of those with the disease who do not get treated is low.

Predictive values can be calculated from sensitivity and specificity as follows:

$$PPV = \frac{Sensitivity}{Sensitivity + (1 - Specificity)}$$

$$NPV = \frac{Specificity}{Specificity + (1 - Sensitivity)}$$

These formulas, however, assume a pretest probability of disease of 50%. Changing the pretest probability of disease will change the predictive values.

PREVALENCE-ADJUSTED PREDICTIVE VALUES

Predictive values have a significant drawback in that they are dependent on the prevalence of the disease in the population studied. A high prevalence will translate to a high PPV. As an extreme example, consider that all patients used in the CRP

study had bacteremia. Because no nonbacteremic patients are included, there is no chance for false-positive test results, and the PPV will be 100%, even if sensitivity is very low. Likewise, a very low prevalence will translate to a low NPV, independent of the test specificity.

Prevalence reflects the prior probability of a disease state. Adjustment of predictive values for prevalence can reduce the dependence of the values on disease prevalence:

$$PPV_{PA} = \frac{Sensitivity \cdot Prevalence}{Sensitivity \cdot Prevalence + (1 - Specificity) \cdot (1 - Prevalence)}$$

$$NPV_{PA} = \frac{Specificity \cdot Prevalence}{Specificity \cdot Prevalence + (1 - Sensitivity) \cdot (1 - Prevalence)}$$

If one were to assume a prevalence of 0.5 (50%), then one can see that these 2 formulas will reduce to the 2 preceding ones.

LIKELIHOOD RATIO AND DIAGNOSTIC ODDS RATIO

The usefulness of a test is further examined by calculation of the *likelihood ratio*. This ratio (for a positive test result) compares the probability of a positive result due to the patient having the disease to the probability of being healthy:

$$LR^{+} = \frac{Positive\ result\ with\ disease}{Positive\ result\ without\ disease}$$

The plus sign indicates that this likelihood ratio is for a positive test result. This likelihood ratio is calculated from sensitivity and specificity:

$$LR^{+} = \frac{Sensitivity}{1 - Specificity}$$

The higher the likelihood ratio for a positive test result, the better the performance of the test for diagnosing the disease (minimizing false-positive results). A value of 10 or greater is considered characteristic of a good test.

If one examines this formula, it is noted that the calculation uses all 4 of the classification cells in the binary classifier in Figure 3. This suggests that prevalence information is implicit in the likelihood ratio, and thus the ratio is less sensitive to prevalence than the PPV and NPV. The likelihood ratio can also be viewed in the context of the pretest probability in that the posttest probability of a disease is derived by adjusting the pretest probability with the test result:

$$Posttest\ Odds = Pretest\ Odds \cdot LR^{+}$$

A likelihood ratio for a negative test result compares the probability of a negative result in a patient with the disease to the probability of being healthy:

$$LR^{-} = \frac{Negative\ result\ with\ disease}{Negative\ result\ without\ disease}$$

A low value reflects better performance, and values less than 0.1 are considered desirable. The value is calculated from sensitivity and specificity as follows:

$$LR^{-} = \frac{1 - Sensitivity}{Specificity}$$

One can combine the likelihood ratios for both positive and negative test results into the *diagnostic odds ratio*:

$$DOR = \frac{LR^{+}}{LR^{-}}$$

This value is perhaps the best single performance evaluator for a given test because it implicitly incorporates the spectrum and prevalence of disease. A value greater than 20 is considered characteristic of a well-performing test.

ROC ANALYSIS

ROC analysis is a method for visualizing classifiers and selecting them based on their performance. The analysis consists of constructing a curve in ROC space that is graphed as the true-positive rate, or sensitivity, plotted against the false-positive rate, or (1 − specificity). The ROC space has several characteristics (Fig 5). The *line of no discrimination* indicates the location in the space where the true-positive rate for a given test (at a given cutoff value) is equal to the false-positive rate, thus unable to discriminate. The triangular area above the line of no discrimination represents the area in which the true-positive rate exceeds the false-positive rate, indicating discriminatory ability. The upper left corner represents a perfectly discriminatory test so that the closer the value to this corner, the better the discriminatory performance. The area below the line of no discrimination indicates a test that predicts incorrectly (false-positive rate greater than true-positive rate). This test mapped to the upper half by inverting the test result (ie, changing a positive result to a negative result and vice versa).

The curve is constructed by varying the cutoff value for classifying a test result as positive from a value below the lowest test result (sensitivity of 100%, specificity of 0%, upper right-hand corner) to a value above the highest test result (sensitivity of 0%, specificity of 100%, lower left-hand corner) as shown in the direction of the arrow in Figure 6. At each cutoff value, the sensitivity and specificity are calculated and the point plotted. The cutoff value that best balances true- and false-positive rates is the one closest to the line that is 90° perpendicular to the no-discrimination line (and hence to the upper left corner). Note that this curve is one in which a positive test result is one above the cutoff value. For a test result in which a positive test result is lower than the cutoff, then the curve is constructed in the opposite direction.

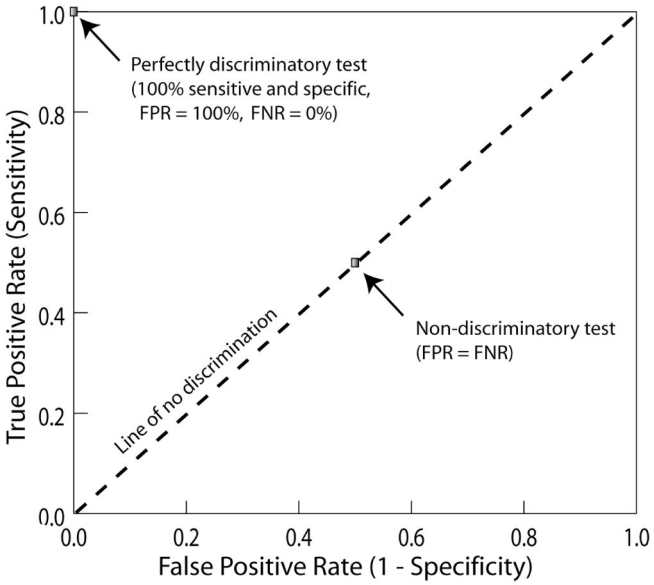


Figure 5. Layout of the receiver operating characteristic space. The true-positive rate (TPR) (sensitivity) is plotted on the dependent axis with the false-positive rate (FPR) (1 − specificity) plotted on the independent axis. The line of no discrimination indicates the location in the space where the FPR for given test (at a given cutoff value) is equal to the TPR, thus unable to discriminate. The triangular area above the line of no discrimination represents the area in which the TPR exceeds the FPR, indicating discriminatory ability. The upper left corner represents a perfectly discriminatory test so that the closer the value to this corner, the better the discriminatory performance. FNR indicates false-negative rate.

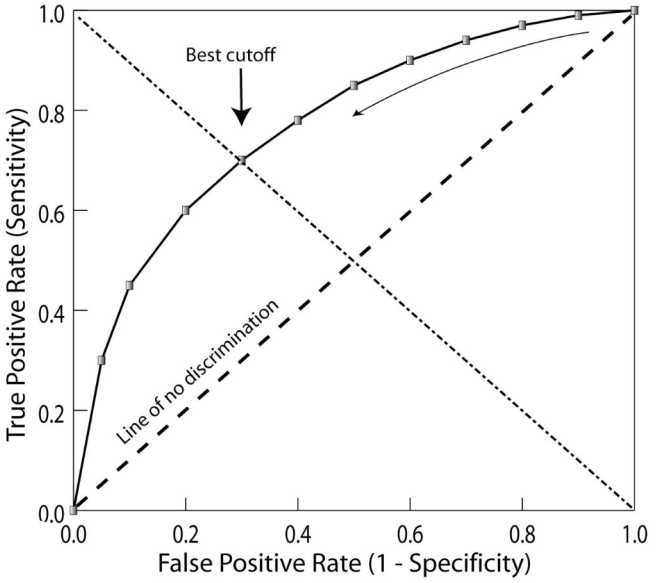


Figure 6. Construction of a receiver operating characteristic curve. The curve is constructed by varying the cutoff value for classifying a test result as positive from a value below the lowest test result (sensitivity of 100%, specificity of 0%, upper right-hand corner) to a value above the highest test result (sensitivity of 0%, specificity of 100%, lower left-hand corner) as shown in the direction of the arrow. At each cutoff, the sensitivity and specificity are calculated and the point clotted. The cutoff value that best balances true- and false-positive rates is the one closest to the line 90° perpendicular to the no-discrimination line (and hence to the upper left corner).

Other information can be derived from the ROC curve. The *area under the curve* (AUC) is a single scalar value that depicts classifier performance. The AUC is the probability that the classifier will rank a randomly chosen positive value higher than a randomly chosen negative value. The higher the value, the better the average performance. The AUC is also equivalent to the probability obtained by performing the Wilcoxon test of ranks on a set of positive and negative test results. Using a single value to characterize the curve, such as the AUC, loses information about how the classifier performs over the range of cutoff values. The *discriminatory index* (d′, d-prime) is a measure that captures both the separation of the curve from the line of no discrimination and its spread. It is calculated as the distance between the mean of activity distribution under noise conditions divided by the SD.

SURVIVAL ANALYSIS

Survival analysis is a branch of statistics that deals with failure in physical and biological systems. Failure in biological systems is usually considered in the context of death, but survival analysis is actually more generalizable to analysis of time to a specified event. Examples in medicine include the analysis of survival after heart transplantation, time to failure of an implantable cardiodefibrillator, and time to development of AIDS after human immunodeficiency virus infection.

Survival analysis involves the description and analysis of survival curves. A *survival curve* describes the survival of a sample over time, typically after an identifiable milestone such as a diagnosis or treatment (Fig 7). A survival function may be described in the following form:

$$S(t) = \Pr(T_i > t)$$

Here, *S(t)* is described as the probability that the time of failure for a given risk factor *i* (*Ti*) is later than time *t*.

The survival curve begins at zero time and continues for the period of observation. Survival at zero time is 1, and the curve is stable or monotonically decreasing over time (which implies that no one comes back to life!):

$$S(t_0) \geq S(t_1), \quad t_1 \geq t_0$$

where *S(t)* denotes the survival at time *t*. The observation period may be sufficiently long so that none of the sample is surviving at the end of the period, but frequently is shorter so that some of the sample is still surviving at the conclusion of the study. The survivors at the end of the observation are denoted *censured* observations because they represent an incompletely observed event. In practice, the period of observation of all of the sample members is not coincident because the initiating events may not be controllable, such as

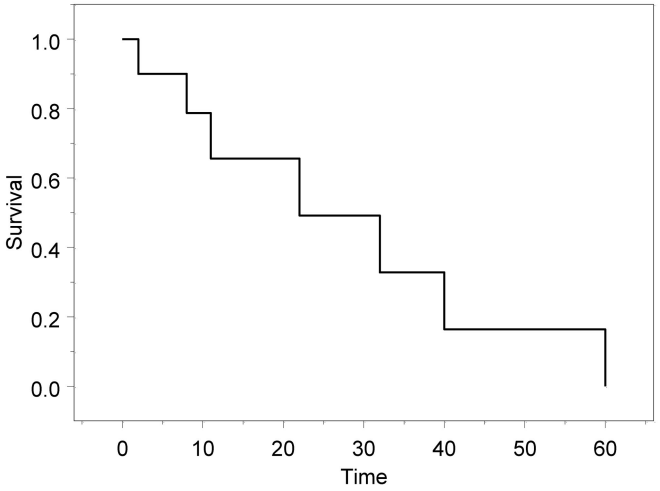


Figure 7. Discrete survival curve graphed as a horizontal step function. Survival is given as a fraction of the starting survival (eg, 1.0). The units on the time axis represent the time units used for recording observation (eg, months for most cancer survival studies).

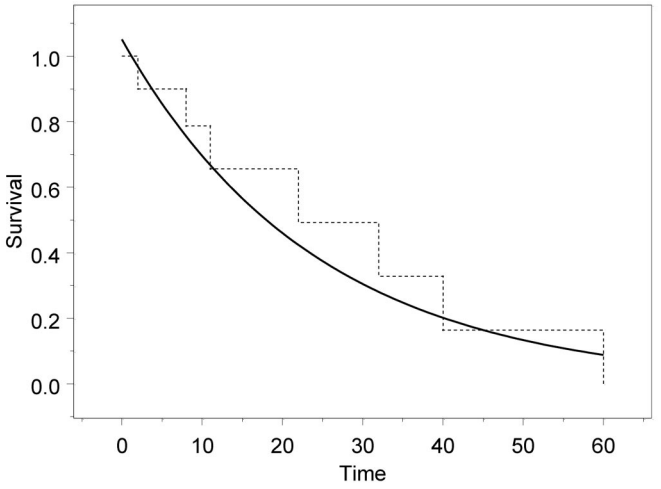


Figure 8. Continuous survival curve of the exponential form (solid line) during 60 months. Shown for comparison (dotted line) is the corresponding discrete survival curve of Figure 7.

the timing of a heart transplantation. Thus, patients may enter into an observation period after the start of the study, and the time of observation may not be the same for all members of the sample. These practical constraints pose challenges to the analysis of survival curves. When the data are presented, the time scale by convention represents the time from entry into the observation period rather than absolute time. This aligns all observation periods and permits graphing the data.

PARAMETRIC SURVIVAL FUNCTIONS

The shape of a typical survival curve is that of a discrete function, with failure or death events recorded as a vertical line indicating the decrease in survival and with horizontal lines between events (Fig 7). As the number of participants observed increases, the discrete pattern gets smaller and in the limit becomes a continuous function that represents the true survival curve for that population and can be fit by parametric models (Fig 8). Several *parametric survival functions* have been used to fit to observed survival data. One of the simplest is the *exponential survival function*:

$$S(t) = e^{-\lambda t}$$

in which λ is the single parameter of the model.

Although most survival curves observed in medicine have an exponential-like form, they are not true exponentials. A more general survival function that better fits most human survival curves is the *Weibull survival function*:

$$S(t) = e^{-(\lambda t)^{\gamma}}$$

The second parameter γ provides an additional degree of freedom for better-fitting observed curves. In most human survival curves, γ is 1 or less. When $\gamma = 1$, we have the special case of the exponential function. Other survival functions, which will not be discussed herein, include the *Gompertz* and *lognormal* functions.

KAPLAN-MEIER ESTIMATOR

The continuous models described herein do not inherently take into account censured data, and dealing with censured observations in these models adds complexity. In medical statistics, the most common approach is to use the nonparametric *Kaplan-Meier estimator*, which can take into account censured data. The Kaplan-Meier estimator $\hat{S}(t)$, also expressed as $\hat{K}$, is the maximum likelihood estimate of the true survival function $S(t)$:

$$\hat{K} = \hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}.$$

In this formula, $\hat{K}$ represents the probability that an individual from the given population will live longer than time t (our observation period). The variable t_i is the observed time of an event (eg, death), n_i is the number at risk when the event occurs (ie, taking into account censured losses), and d_i is the number of deaths occurring at time t_i. An example of the enrollment of patients is depicted in Figure 9. This formula assumes that the values t_i are sorted in ascending order (earliest to latest). It also assumes that censured events are random. The shape of this curve is the stair-step shaped discrete curve as shown in Figure 7. Greenwood developed an estimator of the variance associated with $\hat{K}$ as follows:

$$\text{var}(\hat{K}) = \hat{K}^2 \sum_{t_i < t} \frac{d_i}{n_i (n_i - d_i)}.$$

The calculation of $\hat{K}$ is relatively straightforward.

Table 1 gives sample calculations for a hypothetical set of observations. It shows the step by step calculations to derive

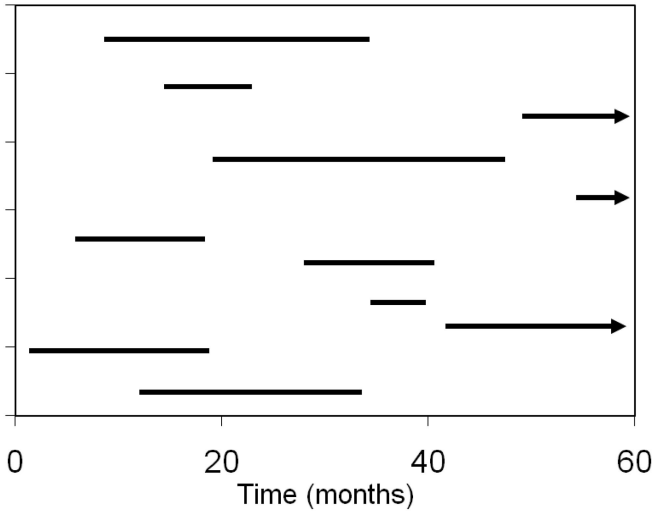


Figure 9. Example of enrollment in a survival analysis trial. The trial runs for 60 months, but patients enter into the trial at various times during the interval (each of which corresponds to time 0 in the analysis). Three patients entered near the end of the trial and were still alive at the end of the observation period. These are considered censured observations because insufficient time was available to completely observe their time course.

the product of individual probabilities. The resulting curve is the one shown in Figure 7.

Mean and median survival times can be calculated from the Kaplan-Meier curve. The *median survival time* is estimated by the end point of the interval containing the survival probability of 0.5. The *mean survival time* is calculated by the sum of the products of the probabilities at the end of each interval and the duration of each interval:

$$\bar{t} = \sum \hat{P}_{i-1}(t_t - t_{t-1})$$

COMPARISON OF SURVIVAL CURVES

There is frequently a need to compare 2 survival curves, for example, to determine the effect of a treatment on survival. Usual regression techniques cannot be used to make this comparison. The underlying distribution typically follows the exponential, Weibull, or other distribution, and not a normal distribution, an assumption required of the usual regression techniques. The introduction of censoring also is not handled by the usual regression techniques. Therefore, regression models that are based on survival curve distributions or are independent of the underlying distribution are required.

Proportional hazards models can be used to compare survival curves. The term *hazard* refers the potential to cause harm (eg, death or other poor outcome). *Hazard rate* refers to the risk of a hazardous outcome. Hazard rates can be obtained from the Kaplan-Meier curve, with the hazard rate at the *i*th interval defined as follows:

$$\hat{\lambda}_i = \frac{\hat{q}_i}{\delta_i} = \frac{d_i}{\delta_1 \cdot n_1}$$

$\hat{q}_i$ = failure probability at i*th* interval

δ_i = length of the i*th* interval

d_i = number of failures in i*th* interval

n_i = number of individuals in the i*th* interval

A *hazard function* expresses these hazard rates as a function of time (λ(t)). A *proportional hazards function* is based on the assumption that a parameter that influences a hazard does so in a multiplicative fashion and is independent of the time. Therefore, a treatment that increases or reduces the hazard does so equally at all time points expressed by the function. One can then express the *effect parameter* c (eg, the parameter that quantitatively describes the effect on the hazard) as the ratio of the 2 hazard functions:

$$\frac{\lambda_1(t)}{\lambda_0(t)} = c \quad or \quad \lambda_1(t) = \lambda_0(t)gc$$

Table 1. Sample calculation of the Kaplan-Meier estimator for a hypothetical survival study

Time, mo	No. at risk	No. censured	No. of failures	$\frac{n_i - d_i}{n_i}$	$\hat{K}$	SD
2	10		1	9/10	0.900	0.095
6	9	1		–	–	–
8	8		1	7/8	0.787	0.134
9	7	1		–	–	–
11	6		1	5/6	0.656	0.163
19	5	1		–	–	–
22	4		1	3/4	0.492	0.187
32	3		1	2/3	0.328	0.183
40	2		1	1/2	0.164	0.147
60	1		1	0	0	NA

Abbreviation: NA, not applicable.

COX PROPORTIONAL HAZARDS MODEL

The *Cox proportional hazards model* does not depend on an underlying distribution (hazard function) and can be considered a semiparametric model. It is therefore applicable to comparing survival curves. The Cox model uses the exponential of a linear function as the effect parameter:

$$\lambda_1(t) = \lambda_0(t)ge^{b_1z_1 + ... + b_mz_m}$$

After a log transformation, the result is a linearized (log-linear) model:

$$\log\left[\frac{\lambda_1(t)}{\lambda_0(t)}\right] = b_1z_1 + ... + b_mz_m$$

in which $\lambda_0(t)$ is the baseline hazard function and $b_1z_1 + ... + b_mz_m$ is the vector of regression coefficients. The part of the equation in brackets is called a hazard ratio. For example, a hazard ratio of 2.0 indicates a double risk of a particular exposure factor compared with the reference exposure.

REFERENCES

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

Requests for reprints should be addressed to:
Steven A. Conrad, MD, PhD
Departments of Medicine, Emergency Medicine, and Pediatrics
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-4228
E-mail: SCONRAD@LSUHSC.EDU

Statistical software programs

Steven A. Conrad, MD, PhD,* and Runhua Shi, MD, PhD†

INTRODUCTION

A number of software applications are available for storage, analysis, and presentation of data, ranging from productivity tools such as Microsoft Excel to comprehensive analysis and reporting packages such as SAS (Statistical Analysis System). Frequently, the tasks of data entry, analysis, and presentation are each performed with separate tools based on the familiarity of the user, particularly when multiple individuals are involved in individual aspects of data handling. It is thus helpful to gain an understanding of the commonly used tools, particularly with respect to their limitations and their ability to exchange data with other applications. With modern computers and powerful software, it is possible to perform analyses that were historically impractical because of computational intensity. This article reviews some of the commonly used programs with the goal of guiding the selection of applications for use in projects. Emphasis will be placed on data types, entry, and exchange, as well as some of the major features of several individual applications.

DATA TYPES

All data analysis applications assume data intended for analysis is in one of several predefined formats. Before any data collection is initiated, the experimental design should be reviewed so that the data entry application can be set up for the required data types. Some applications, such as Excel, by default are very forgiving with data entry, allowing mixed data types that the application attempts to interpret. This can lead to incompatibilities within data sets that can impair data exchange or even worse can result in analyses that are silently inaccurate. This section provides an overview of computational data types (how data are stored and handled by computer systems) and statistical data types (how data are viewed for analysis by statistical software). Further guidelines will be provided under the discussion of individual software applications.

Computational Data Types

Computer software represents and operates on numbers in 1 of 2 basic formats: integer and floating point. *Integer* values, as the name suggests, have no fractional component and take on only integral values of 0, 1, 2, 3, etc. The maximum integer value available is dependent on the computer architecture, with modern systems accommodating the ranges of $\pm 2^{31}$ ($\pm$2,147,483,648) for 32-bit systems or $\pm 2^{63}$ ($\pm$223,372,036,854,775,808) for 64-bit systems. It is unlikely that any analysis problem will ever approach these limits. *Floating point* includes the storage of numbers in a format similar to scientific notation, with both a number component and an exponent. The term *floating point* indicates that there is a constant number of significant digits as a result of "floating" the decimal point while correspondingly altering the exponent. Current systems use the floating point standards established by the Institute of Electrical and Electronics Engineers. The precision of the number component and the range of the exponent vary with the computer architecture, with *single precision* having approximately 7 digits of precision with exponents ranging from 10^{-38} to 10^{+38}. Although data observations will only rarely exceed this range, the limited precision can be problematic when intermediate values are generated during computations. Double precision floating point handles approximately 17 digits of precision and an exponent range of 10^{-308} to 10^{+308}, assuring both improved accuracy and precision of intermediate results during computation. Fortunately, floating point computations on essentially all computer systems default to double precision so that entry and storage of data in single precision floating point format can save space without any computational penalties. Some vendors provide software emulation of larger (128 bit) numbers such as the *decimal* type, which have a much higher precision (approximately 29 digits) but lower exponent range (10^{-28} to 10^{+28}). It should be recognized that floating point computations technically are only approximate due to the fixed precision, but the number of digits in double precision format makes this approximation of little practical consequence in most cases.

Dates and times in software applications are represented internally as a *date-time* type, that is, a binary interpretation of the date and time. Unlike floating point numbers, there is no widely accepted standard for storage of date-time values, making data interchange among systems difficult. This is usually accomplished through the conversion to a standard character format such as "2006–08–23 15:45" that can be produced and interpreted across systems. With software programs, such as SAS or SPSS, that have explicit date data types, the variables can be manipulated arithmetically to calculate time elapsed between dates, for example.

Statistical Data Types

Statistical analysis applications view data in 1 of 3 contexts. *Observational data types* include data that typically represent the dependent variables in an experimental plan. Parametric analyses expect floating point values. Integer values can be

Affiliations: *Department of Medicine, Louisiana State University Health Sciences Center, Shreveport, Louisiana; †Department of Medicine, Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Louisiana.

Disclosures: Authors have nothing to disclose.

Received for publication March 11, 2009; Received in revised form April 22, 2009; Accepted for publication March 25, 2009.

substituted when they represent an appropriate substitute for floating point values. Nonparametric analyses can use either type, but integer variables should be used if the data represent true integral values. Programs that default to floating point values will require explicit designation of these variables as integers.

Classification data types are used in statistical models, such as analysis of variance and logistical regression. These data typically represent nominal or ordinal factor levels that are predetermined in the experimental design. Formal statistical programs, such as SAS and S-Plus, allow the designation of variables as factors. Traditionally, these data are often coded in numerical format (eg, 1 for male and 2 for female), but these programs allow the use of descriptive values (eg, male or female) that make data management far more intuitive and realistic. If the application supports descriptive factor levels or can import character strings representing factor levels, then taking advantage of this descriptive feature is advisable. Applications targeted for data entry, such as Access or Excel, can take advantage of features such as supporting data tables or automemorized dropdown lists, respectively.

Descriptive data types serve to annotate the values of variables and factors to facilitate interpretation of presentation data and graphs but are not variables analyzed directly. This type is represented by a textual data element that may be included as a property of the underlying variable or factor or assigned during reporting or graphing (eg, male = 1, female = 2).

MICROSOFT EXCEL AND OTHER SPREADSHEETS

Microsoft Excel is a productivity tool known as a *spreadsheet*, providing a grid of cells in which each cell can hold data values or computational formulas referencing other cells. The first commercially successful spreadsheet for personal computers (Apple) was Visicalc. Lotus 1–2-3 was a successful spreadsheet that ran on the IBM personal computer. Microsoft Excel competed with Lotus 1–2-3 and has became the currently dominant spreadsheet. Macro languages now extend spreadsheets by providing a programming infrastructure.

Because of its column and row grid orientation, Excel has become a popular platform for recording experimental data. Its computational capabilities also permit simple summary statistical calculations using built-in formulas (eg, AVGERAGE or STDEV) or more extensive statistical calculations using add-on packages (eg, Analysis ToolPak). A principal advantage of Excel is that it is a popular application distributed with the Microsoft Office suite, widely available on Windows and Mac platforms. The file format is interchangeable between these 2 platforms.

Excel has a number of limitations that users should consider when choosing an application for statistical analysis. It has an unstructured interface. The user has to indicate the purpose and role of each set of data cells in calculations, leading to potential for error. It is not suitable for large data sets. The calculation formulas are interpreted cell by cell and are optimized for flexibility but not performance. The types of data that can be stored are limited, consisting of floating point numbers, date-time values, and character strings. It does not support true integer storage. Excel is not well suited for complex data sets, such as hierarchical data, that may require analysis at several levels (a very common scenario!). Storage of hierarchical data requires redundancy that can lead to errors occurring during data entry. A better data storage platform for this purpose would be a relational database application such as Microsoft Access or a database server (eg, Oracle or SQL Server). Spreadsheets are not well suited for complex analyses. Also, they are not easily adapted to data that require analysis by groups. A number of graph formats are provided, but customization of individual graphs is limited, and thus it is not well suited to generating publication quality graphs except in simple scenarios.

Suggestions for using Microsoft Excel are as follows:

- Avoid Excel's attempt at data type interpretation by explicitly assigning an appropriate format to each column of data.
- Define integer values as numbers without decimal points or fractions through Excel's cell formatting capabilities.
- Do not separate date and time values into separate columns if both components are required. Use a single column that defines both and time.
- For text fields it may be important to have separate columns for subfields, such as last name, first name rather than a single name field. If it were necessary to search for someone by last name, for example, this is a difficult task if first and last names are in the same column.
- Keep data columns for data only and resist placing text notes or computational results in the same columns as data.
- Keep the data type consistent within each column. This will facilitate exportation to a statistical package if needed and avoid errors in analysis.
- Leave a cell blank if that data point is missing. If data are to be exported to a statistical analysis program, that program may allow predefined values to be interpreted as missing values on import.
- Keep column names short so that they are easy to manipulate when imported into analysis or presentation software. Some packages do not allow spaces in column names, so consider removing spaces or replacing with a character such as an underscore.

In summary, Excel is perhaps best suited as a data entry tool for simple experimental designs, either for exporting to formal statistical packages or for performing exploratory analyses.

MICROSOFT ACCESS

Access is a file-based relational database application that is well suited for data entry, structured data organization, and flexible data retrieval. As a database application, a wide variety of data types can be stored, including several variants

of integers, floating point data with different levels of precision, date-time data, character data, and fixed and variable-length textual string data. Through its relational structure, supporting tables can be used to both offer available choices for classification variables and ensure that only appropriate data can be entered. Complex data organizational structures can be imposed to support data integrity. Its query capabilities allow for flexible retrieval of hierarchical data into a flat format suitable for a statistical analysis program. These queries can be stored for ease of subsequent data retrieval as the database increases in size.

Access is based on industry-standard protocols that enable many statistical analysis programs to use its data directly for analysis without having to first import it. Data entry is facilitated by the creating of data entry forms that guide data entry and validate it. Because it is file based, the database file can be transferred easily among computers, but its multiuser capabilities allow users on different computers to enter or retrieve data from a common shared database file, eliminating the need to move the file around to different users.

Access is also part of some Microsoft Office suites and thus is widely available. It does not offer any statistical analysis capabilities, so it is only suitable for data entry and reporting. However, it offers data management capabilities not available in statistical analysis applications and deserves strong consideration as a data management tool to partner with these applications.

Client-server databases, such as Oracle and Microsoft SQL Server, offer the data handling features of Access but geared for high-availability multiuser access, such as with an online Web-based database. Data can be accessed through industry-standard protocols such as ODBC, but these systems do not offer the forms capabilities of Access. It is possible, however, to use Access forms that connect to a database server for a more custom solution.

SAS SYSTEM

The SAS System (previously known as Statistical Analysis System) is a comprehensive software package available for data handling, reporting, and statistical analysis. It was originally developed for mainframe computers but has since migrated to essentially all available platforms. It couples a procedural programming language to a fourth-generation language based on an extensive library of data handling and analysis functions. Graphical user interfaces (GUIs) have been developed to assist users who want to avoid a programming language construct, but the programming language approach remains the choice of many if not most SAS users. The ability to enter data into a dataset using a spreadsheet format is supported. One of the authors has used SAS on a desktop computer to perform 30,000 analysis of variance (ANOVA) calculations of a 3-factor linear model over a total of 12 million gene expression values with only 6 lines of SAS code and a solution time of 14 seconds.

SAS is distributed as a set of approximately 30 modules that add functionality to its core software. The core software (Base SAS) provides data handling and reporting and basic summary statistical analyses. Perhaps the most commonly used module is SAS/STAT, which provides a large number of commonly used statistical procedures. Some other examples of the many modules include SAS/INSIGHT for exploratory data analysis, SAS/IML for matrix manipulation, SAS/GRAPH for presentation graphics, and SAS/QC for quality improvement procedures.

Analyses in SAS are performed on *datasets*, which can be created programmatically or by importing from other programs. A simple data step example program to create a dataset named ONE from inline data values is as follows:

```
DATA ONE;
INPUT DRUG $ SBP;
DATALINES;
A 140
A 128. . .
B 135
B 156. . . ;
RUN;
```

where A and B designate 1 of 2 treatments and the 3-digit numbers represent systolic blood pressure (SBP).

Using the SAS Import procedure to import and generate a dataset named ONE from a sheet in an Excel workbook is as follows:

```
PROC IMPORT DATAFILE= "C:\Data.xls" OUT=ONE;
SHEET="Sheet1";
GETNAMES=YES;
RUN;
```

Different versions of SAS accept only specific versions of the MS Excel spreadsheet and generally lag somewhat behind in acceptance of new formats. For example, version 9.1.3 (most recent as of this writing) will accept Microsoft Excel 2003 format but not 2007 format. Verify compatibility before saving the spreadsheet. Because Excel allows multiple sheets within a file, the sheet name will have to be specified as indicated in the above code example.

Familiarity with how SAS expects data in its datasets will facilitate data importing when data from other applications such as Excel are imported into SAS. Most SAS procedures are built on the premise that each line in the dataset represents a single observation. Each observation may consist of more than one observed variable, but each variable appears in only one column. The observation could also include other variables, such as factor levels and grouping variables. For example, data consisting of SBPs from 3 groups to be compared by an ANOVA would look like the following:

```
Drug SBP
A 140
A 128
. . .
B 135
B 156
. . .
C 122
C 118
```

. . .

Notice that the observed variable SBP appears only once per entry. This is distinct from how users usually record data in a program such as Excel:

SBP_A SBP_B SBP_C
140 135 122
128 156 118
. . .

In this case, Excel expects the data to be in 2 or more columns that will be compared. In SAS, the GROUP variable is provided to the ANOVA procedure to indicate the treatment group. To move this latter dataset structure into SAS, the Excel data would either have to be restructured to the format in the first dataset, or a SAS data step would have to be created in which the 3 data values per row from dataset ONE were exported to 3 separate observations into a new dataset TWO as follows:

```
DATA TWO; SET ONE;
Drug = 'A'; SBP = SBP_A; OUTPUT;
Drug = 'B'; SBP = SBP_B; OUTPUT;
Drug = 'C'; SBP = SBP_C; OUTPUT;
KEEP Drug SBP;
```

Statistical analyses are performed by the specification of 1 or more procedure statements that reference a dataset. A SAS program to perform a balanced 1-way ANOVA follows, where the CLASS statement identifies the classification variable:

```
PROC ANOVA DATA=ONE;
CLASS Drug;
MODEL SBP = Drug;
```

The MODEL statement is the means of specifying statistical models to all of the SAS procedures that analyze linear and nonlinear models (eg, ANOVA, REG, GLM, and others). A MODEL statement that specifies a 2-way ANOVA with interactions is as follows:

```
MODEL SBP = Drug Condition Drug*Condition;
```

or equivalently:

```
MODEL SBP = Drug|Condition;
```

The original SAS graphing capabilities were designed for line printers (SAS preceded modern output devices!). The following procedure will produce a character-based plot of these data:

```
PROC PLOT;
PLOT SBP*Drug;
```

Replacing PLOT with GPLOT will produce a high-resolution graph to the default output device (typically the display monitor).

The advantages of SAS are numerous, so only a few will be highlighted. The data manipulation capabilities are exhaustive and supported by a 4GL language based on a comprehensive procedure library. SAS works with data directly on disk; thus, it can handle datasets as large as the disk capacity on the user's computer or network drive. The code has been extensively verified and optimized during the past 40 or more years, so it is computationally fast, yielding trustworthy results. The data analysis procedures are among the most extensive available, offering analyses not available in most other packages. It is particularly well suited when data analysis is a recurring task because the SAS programs that are developed can be easily invoked on new data.

SAS requires an initial investment of time to learn, and it is not as intuitive as many other packages. Its graphing capabilities are also somewhat more limited than many other statistical packages.

In SAS interactive GUI mode, all basic and many comprehensive statistical analyses can be performed, in addition to graphing and reporting. However, the ability to access all procedures requires use of the programming language.

SPSS (STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES)

SPSS was developed at Stanford University in 1968 and further developed at the University of Chicago. It was so widely received that it quickly became a small business and had to separate from the university and become incorporated in 1975. Although developed initially for social scientists, the statistical capabilities of SPSS and its relative user-friendliness led to its adoption by statisticians and scientists practicing in many fields. Like SAS, SPSS has a command language syntax that allows programs to be written and saved for later use. Initial implementations were on large mainframe computers. For many years, the command language was the only way to use SPSS. The advent of personal computers dramatically changed the way statistical analysis was performed. In the mid-1980s, SPSS was the first major statistical package to be ported to personal computers; it ran under DOS, the original Microsoft operating system, and still used the command language interface. In 1992, SPSS was released in a Windows compatible format and became the first major statistical package to have a GUI. Today SPSS is available for most operating systems, including Windows, Macintosh, and Linux.

SPSS retains its command language interface capabilities, but most of its features can be used in an interactive graphical interface using extensive pull-down menus. For the interactive use, data are entered through a spreadsheet-like interface or imported from a variety of source files, such as Excel, ASCII text files, and databases via ODBC and SQL. The data entry screen has 2 options, data and variable modes. Variable mode allows the addition of labels, missing values, and data type specification (eg, date, string, categorical or floating point). Data mode displays the data in a spreadsheet format. A number of sorting, aggregation, case selection, and data transformation functions are available. Data can be saved in a special SPSS format containing all the data entry, labels, transformations, and other defined characteristics, which greatly simplify future analysis. Once data entry and data manipulations are complete, statistical procedures can be selected through a pull-down menu. For each statistical procedure chosen, a window is shown that allows variable selection and specification of various options particular to the chosen analysis. Output is immediately shown on the com-

puter screen and can be saved in a text format for future reference. SPSS also contains a powerful graphing capability that can be used with either command language or graphical interface. The graphs, usually after some customizing, are publication quality and can be exported or cut and pasted into other applications. Some of the graphs for this series of articles were created with SPSS.

An example of an ANOVA calculation in the command language follows:

```
UNIANOVA FVC BY Smoke Race WITH Age BMI
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/CRITERIA=ALPHA(0.05)
/DESIGN=Age BMI Smoke Race Smoke*Race.
```

BMI indicates body mass index, and FVC indicates forced vital capacity. Although this is straightforward for the experienced statistician, the typical user would have difficulty remembering the commands, options, and syntax. The scripts corresponding to the graphical commands can be saved for duplication or modification at a future time.

For the typical user, the number and size of records that can be processed with SPSS are virtually unlimited, restricted only by the hard disk, rapid access memory, and speed of their processor. One of the authors has processed analyses with tens of thousands of large records on a Windows personal computer without problem. SPSS comes in several modules. The SPSS base module contains the most commonly used statistical tools; specialized modules can be added for different applications, such as classification trees or neural networks. Student versions are relatively inexpensive and have the same capabilities as the full SPSS has but with limitations on the number of cases that can be read (1,500). For most students, this is not a serious limitation.

In summary, SPSS is a powerful, mature statistical package that is easy to use. Although SAS is, arguably, the most widely used by professional statisticians and data analysts, SPSS at this time has a much more attractive and easy to use graphical interface. There are few common statistical procedures and data manipulation capabilities not available in SPSS. Of course, there is always the danger of easy-to-use statistical programs, not limited to SPSS, for which naive users may use the wrong analysis or misinterpret the results.

S-PLUS, 'S,' AND 'R'

The S-Plus statistical analysis package is a comprehensive application based on the 'S' language. This language is a functional and procedural programming language that incorporates an extensive library of several thousand functions for performing data manipulation, mathematical computations, statistical analysis, and graphing and is well suited to scientific and engineering computations. Complex data objects can be created and manipulated in S-Plus, and it provides support for matrix and vector operations, including symbolic processing, such as those seen in mathematical packages such as MATLAB. 'S' now fully supports an object-oriented programming paradigm, a feature not available in most statistical packages. The software is available on common operating system platforms (Windows, Macintosh, and Linux)

S-Plus has a powerful user interface that provides an intuitive approach to data management and statistical analysis. As a result, 1-time analyses can be performed without the need to program in the 'S' language. Data are stored in datasets much like SAS, but some functions expect multiple observations in each data row. Data can be imported from a variety of sources or can be directly entered into the dataset in a spreadsheet-like manner. Functions to transform and reshape data are provided. Earlier versions operated on data residing in memory that limited the size of datasets, but current versions have large dataset capabilities by sharing data between memory and disk. S-Plus supports a number of data types, including factors, character strings, integers, single and double precision floating point, logical, date-time, and complex numbers.

The graphing capabilities in S-plus are extensive and support interactive modification. For example, graphs can be easily annotated, the axes can be moved by dragging, and plot types can be easily changed. It is simple to create multiple plots per graph and to have multiple graphs per page.

The 'S' language was developed by Bell Laboratories (now Lucent Technologies) in the 1970s and commercially licensed since the 1980s. It is currently licensed by Insightful Corporation (now TIBCO Software Inc) as part of the S-Plus package. It is an open language in that it can be easily extended with additional functions and existing functions can be modified. The 'R' language is an open source implementation of the 'S' language licensed under the GNU General Public License and is largely compatible with the 'S' language. Being open source, it is widely used, but support is not available except through public forums. S-Plus can now import packages written in 'R'.

The 'S' language follows a compact functional notation. Sample 'S' code to import data from an Excel spreadsheet and perform an ANOVA follows (assumes Excel has the data as a single observation per row with columns named SBP and Drug):

```
> dataone <- importData("a:\\examples\\bp.dat")
> aov(formula = SBP ~ Drug, data = dataone)
```

Numerous options to the functions performing analyses are available, but they default to commonly used values and thus need not be included unless different options are needed. S-Plus uses a different notation than SAS for specifying statistical models. In this example, the model specification SBP ~ Drug indicates that SBP is the dependent variable and Drug the single independent variable. Polynomial regression (quadratic) would be specified as SBP ~ Drug + Drug^2.

A model with main effects and interaction would be specified as SBP ~ Drug + Condition + Drug:Condition, or equivalently SBP ~ Drug * Condition. Because S is a functional language, function arguments can be replaced with functions, allowing this to be expressed as a compact, single-line program:

```
> aov(formula = SBP ~ Drug, data = importData("a:
\\examples\\bp.dat"))
```

S-Plus is popular because of its refined user interface and interactive graphing capabilities and is especially well suited for scientific data analysis because it incorporates extensive support for scientific functions and calculations. Several add-in modules are available for performing microarray data analysis, wavelet and signal analysis, environmental statistics, and optimization, among others. Using S-Plus from the GUI is relatively intuitive and supports most statistical tasks. Learning the 'S' language for more advanced analysis requires some investment in time.

CONCLUSION

In addition to a working knowledge of the statistical tests to be used, optimal use of data management and statistical software requires an understanding of how data are stored and manipulated, as well as the strengths and weaknesses of the analysis software in use. Each statistical software package offers advantages and disadvantages, and it is not uncommon to use more than one package for data manipulation and analysis. Most statistical analysis packages can also be used to produce tables and graphs in addition to performing analyses. Whatever software package is used, an investment of time in learning the details of its use will yield many returns, reduce errors, and improve productivity.

REFERENCES

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

Requests for reprints should be addressed to:
Steven A. Conrad, MD, PhD
Department of Medicine
Louisiana State University Health Sciences Center
1501 Kings Highway
Shreveport, LA 71103-4228
E-mail: SCONRAD@LSUHSC.EDU

REFERENCES

Agresti A. *Categorical Data Analysis.* New York, NY: Wiley; 2002.

Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ.* 1994;308:1552.

Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ.* 1994;309:102.

Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research.* 4th ed. Maiden, MA: Blackwell Science; 2002.

Aschengrau A, Seage GR III. *Essentials of Epidemiology in Public Health.* 2nd ed. Sudbury, MA: Jones and Bartlett Publishers; 2008.

Bryant TN. Presenting graphical presentation. *Pediatr Allergy Immunol*. 1999;10:4–13.

Bryant TN. The presentation of statistics. *Pediatr Allergy Immunol*. 1998;9:108–115.

Casella G, Berger RL. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Thomson Learning; 2002.

Cleveland WS. *The Elements of Graphic Data.* New Jersey: Hobert Press; 1994.

Cody RP, Smith JK. *Applied Statistics and the SAS Programming Language.* 5th ed. Upper Saddle River, NJ: Prentice Hall; 2006.

Comprehensive Meta Analysis [software program]. Englewood, NJ: Biostat; 2008.

Daintith J. *A Dictionary of Computing.* New York, NY: Oxford University Press; 2004.

Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences.* 8th ed. Atlanta: Georgia State University; 2005.

De Vet H. Observer reliability and agreement. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics.* 2nd ed. Chichester, England: Wiley; 2005:3801–3805.

Everitt B, Rabe-Hesketh S. *Analyzing Medical Data Using S-PLUS.* New York, NY: Springer-Verlag; 2001.

Everitt BS. *A Handbook of Statistical Analyses Using S-PLUS.* 2nd ed. Boca Raton, FL: CRC Press; 2002.

Fawcett T. An introduction to ROC analysis. *Pattern Recog Lett*. 2006;27:861–874.

Green SB, Salkind NJ. *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data.* 5th ed. Upper Saddle River, NJ: Prentice Hall; 2008.

Hankinson SE, Willett WC, Colditz GA, et al. Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *Lancet*. 1998;351:1393–1396.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.

Hedges L, Olkin I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.

Hedges LV. Fixed effects models. In: Hedges H, Ca LV, eds. *The Handbook of Research Synthesis.* New York, NY: Russell Sage Foundation; 1994:285–299.

Hulley SB, Cummings SR, eds. *Designing Clinical Research.* Baltimore, MD: Williams & Wilkins; 1988.

Jekel JF, Katz DL, Elmore JO, Wild DMG. *Epidemiology, Biostatistics and Preventive Medicine.* 3rd ed. Philadelphia, PA: Saunders Elsvier; 2007.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457–481.

Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression.* 3rd ed. Pacific Grove, CA: Duxbury Applied; 2007.

Krause A, Olson M. *The Basics of S-Plus.* 4th ed. New York, NY: Springer-Verlag; 2005.

Matthews DE, Farewell VT. *Using and Understanding Medical Statistics.* 4th ed. New York, NY: Karger; 2007.

Maxwell SE, Kelley K, Rausch JR. Sample size planning statistical power and accuracy in parameter estimation. *Annu Rev Psychol*. 2008;59:537–563.

Mcneil D. *Epidemiological Research Methods.* New York, NY: Wiley; 1996.

Motulsky H. *Intuitive Biostatistics.* New York, NY: Oxford University Press; 1995.

Norman G, Streiner L. *Biostatistics: The Bare Essentials.* 3rd ed. Hamilton, Ontario: BC Decker Inc; 2008.

Peng CYJ. *Data Analysis Using SAS.* Thousand Oaks, CA: SAGE Publications; 2008.

Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford, England: Oxford University Press; 2003.

Pocock SJ. *Clinical Trials: A Practical Approach.* Chichester, England: Wiley; 1983.

Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926–930.

Riegelman RK. *Studying a Study and Testing a Test: How to Read the Medical Literature*. Boston, MA: Little Brown & Co; 1981.

Rosner B. *Fundamentals of Biostatistics.* 6th ed. Pacific Grove, CA: Duxbury; 2006.

Sacks H, Chalmers TS, Smith H. Randomized versus historical controls for clinical trials. *Am J Med*. 1982;72:233–240.

SAS/STAT 9.1 User's Guide. Vol 1–7. Cary, NC: SAS Publishing; 2004.

Silman AJ, Macfarlane GJ. Which type of study? In: *Epidemiological Studies: A Practical Guide.* 2nd ed. Cambridge, England: Cambridge University Press; 2002:31–41.

Sprinthall RC. *SPSS From A to Z: A Brief Step-by-Step Manual.* Boston, MA: Allyn & Bacon; 2009.

Software program for graphs: SigmaPlot version 11. San Jose, CA: Systat Software Inc. Available at: www.sigrnaplot.conilproducts/SigmaPlot/.

Valiela I. *Doing Science: Design, Analysis, and Communication of Scientific Research.* New York, NY: Oxford University Press; 2001.

Van Belle O. *Statistical Rules of Thumb.* 2nd ed. New York, NY: John Wiley & Sons; 2008.

Wainer H. How to display the data badly. *Am Stat*. 1984; 38:137–147.

Wallgren A, Waligren B, Persson R, et al. *Graphic Statistics and Data.* Thousand Oaks, CA: Sage Publications; 1996.

Weisberg S. *Applied Linear Regression.* 3rd ed. Hoboken, NJ: Wiley/Interscience; 2005.

Yu H, Jin F, Shu XO, et al. Insulin-like growth factors and breast cancer risk in Chinese women. *Cancer Epidemiol Biomarkers Prev*. 2002;11:705–712.

Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561–577.

Glossary of common statistics terms

ANCOVA – analysis of covariance. ANOVA with 1 or more continuous covariates in addition to factors. The analysis adjusts for significant covariates before testing for factor (group) effects.

ANOVA – analysis of variance. A statistical test to compare the means of several groups simultaneously. It assumes random sampling from independent normally distributed data. Multiple factors (eg, treatments, groups, etc) can be considered simultaneously. ANOVA generalizes the Student *t* test to more than 2 groups.

Bonferroni correction – a simple, but conservative, method of adjusting *p* values in multiple comparisons situations. The idea is to divide *P* by the number of planned comparisons to get a smaller *p* value. This eliminates the problem of increased type I error (ie, testing until you find something significant).

Critical value – a preset value of a test statistic (ie, critical value is 1.96) that reflects the level of probability used for rejecting the null hypothesis. The value determines the probability of a type I error, incorrectly rejecting the null hypothesis, that the investigator is willing to accept. This probability is traditionally called α.

Degrees of freedom – the number of independent observations associated with an estimate of variance or of a component of an ANOVA is free to vary. For example, in an ANOVA with 3 groups (namely 1, 2, and 3), with a sample of n 1 observations, the degrees of freedom is $n - 1$ when you calculate the mean for this group; 1 *df* is used in calculating the mean value from which the variance can be calculated.

Interval level of measurement – requires that data can be ordered (unlike nominal data) but also that the distance between levels is the same. For example, degrees Celsius: the distance between 10° and 15° is the same as the distance between 15° and 20°, for example. Subtraction and addition are meaningful; however, choice of a zero point is arbitrary.

Kruskal-Wallis test – a nonparametric test to compare more than 2 groups simultaneously. It is an extension of the Mann-Whitney 2-sample test.

Mann-Whitney – a nonparametric method for comparing 2 groups. It is based on ranking the raw data and comparing mean ranks between groups.

Multiple comparisons problem – the inflation of type I error (the probability of rejecting the null hypothesis when it is true) that occurs when multiple statistical tests are performed. The more tests that are performed the more likely it is that a test will be declared significant by chance alone when it isn't significant.

Normal distribution – an important probability distribution completely specified by a mean and SD. The data follow a familiar symmetric bell-shaped distribution. The normal distribution is also called a gaussian distribution because it was first described mathematically by Carl Friedrich Gauss.

Null hypothesis – the hypothesis that assumes no effect of the experiment (ie, no difference between outcome variables for the 2 or more groups being studied).

Parameter – a predefined measurement on a population that is used to characterize a feature of the population. Examples include the population mean and SD.

Population – the complete collection of items or subjects of interest in a study that share a common feature to be measured. A population is usually too large or inaccessible to enable observations on every subject, thus the need for sampling. An example of a population might be all patients with hypertension in a given geographic region.

Power – the probability of correctly rejecting the null hypothesis (eg, concluding that 2 treatments are different when in fact they are). Power is often expressed as a percentage. For study planning, power levels of 80%, 90%, and 95% are commonly used. The power is equal to $1 - \beta$ (see type II error).

P value – the probability of incorrectly rejecting the null hypothesis. When *p* is less than a prespecified number, the critical value, the null hypothesis is rejected. Traditionally, $p < .05$ is grounds for rejecting the null hypothesis, but other values can be used at the judgment of the investigator, depending on the situation.

Random sampling – a method of selecting subjects in a manner that eliminates bias. Random number tables generated by computational methods are commonly used. There are 5 commonly used methods of random sampling: simple random sampling, systematic sampling, stratified sampling, cluster sampling, and multistage sampling.

Repeated-measures ANOVA – a special form of ANOVA in which 1 or more of the factors is measured repeatedly on each subject. Because the repeated measures are correlated, special methods of computing mean square errors are necessary.

Sample – a subset of items or subjects drawn from a population to draw inferences about the population. The most common approach to sampling is random sampling.

SD – standard deviation, a measure of variability in a sample, calculated from the squared distance from the mean for each value and the sample size in a sample.

SE – standard error of the mean, calculated as follows: $SE = SD/\sqrt{n}$, where S is the SD and n is the sample size.

Statistic – a predefined measure obtained on a sample that characterizes the sample, obtained for the purposes of estimating that measure in the population (see parameter). Examples include the sample mean and sample SD.

Statistically significant – a study conclusion based on a p value less than the prespecified critical value is considered to be statistically significant. This means that the probability of rejecting the null hypothesis by chance is acceptably low. Commonly this means $p < .05$, but other values may be used.

Student t distribution – a probability distribution similar to the normal distribution that is used to estimate the mean of the normal distribution when the sample size is small. The width of the distribution is dependent on the size of the sample and approaches the normal distribution when the sample size gets large.

t test – a test that uses the t statistic with the Student t distribution to test hypotheses about the mean of a population or for comparing the means of 2 populations.

Type I error – rejecting the null hypothesis when it is actually correct (eg, saying a new drug is better than the standard treatment when in fact the 2 drugs are equivalent). The probability of a type I error is designated as α.

Type II error – failing to reject the null hypothesis when it is not true (eg, saying 2 drugs are the same when in fact they are different). The probability of a type II error is designated as β.

$\bar{X}$ (pronounced "x bar") – the usual symbol for the mean of a sample.