The Laryngoscope
Lippincott Williams & Wilkins, Inc.

Tutorials in Clinical Research: Part VII. Understanding Comparative Statistics (Contrast)–Part A: General Concepts of Statistical Significance

J. Gail Neely, MD; James M. Hartman, MD; James W. Forsen, Jr., MD; Mark S. Wallace, MD

Objectives/Hypothesis: **The present tutorial is the seventh in a series of Tutorials in Clinical Research. The specific purpose of the tutorial (Part A) and its sequel (Part B) is to introduce and explain three commonly used statistical tools for assessing contrast in the comparison between two groups.** ***Study Design:*** **Tutorial.** ***Methods:*** **The authors met weekly for 10 months discussing clinical research studies and the applied statistics. The difficulty was not in the material but in the effort to make the report easy to read and as short as possible.** ***Results:*** **The tutorial is organized into two parts. Part A, which is the present report, focuses on the fundamental concepts of the null hypothesis and comparative statistical significance. The sequel, Part B, discusses the application of three common statistical indexes of contrast, the χ^2, Mann-Whitney *U*, and Student *t* tests.** ***Conclusions:*** **Assessing the validity of medical studies requires a working knowledge of research design and statistics; obtaining this knowledge need not be beyond the ability of the busy surgeon. The authors have tried to construct an accurate, easy-to-read, easy-to-apply, basic introduction to comparing two groups. The long-term goal of the present tutorial and others in the series is to facilitate basic understanding of clinical research, thereby stimulating reading of some of the numerous well-written research design and statistical texts. This knowledge may then be applied to the continuing educational review of the literature and the systematic prospective analysis of individual practices.** ***Key Words:*** **Research methods, research design, clinical protocols, statistics methods.**

Laryngoscope, **113:1534–1540, 2003**

From the Clinical Research Working Group, Department of Otolaryngology—Head and Neck Surgery, Washington University School of Medicine, St. Louis, Missouri, U.S.A.

This Manuscript was accepted for publication May 9, 2003.

Send Correspondence to J. Gail Neely, MD, Department of Otolaryngology—Head and Neck Surgery, Washington University School of Medicine, 660 South Euclid Avenue, Box 8115, St. Louis, Missouri 63110, U.S.A. E-mail: jgneely@aol.com

INTRODUCTION

A previous tutorial, "Part VI: Descriptive Statistics," focused on *univariate analysis*, a process in which the "spectrum," or distribution, of a single variable is described.[1] The present tutorial is the first of two tutorials concerned with *bivariate analysis*, a process in which the relationship between two variables is explored.[2] One variable is the *independent variable* (sometimes called the predictor variable, and often noted as "X" and placed on the x-axis of scatter plots), and the other is the *dependent variable* (also known as the outcome variable, and often noted as "Y" and placed on the y-axis of scatter plots).[3,4]

The focus of the two tutorials (Parts A and B) is on the *comparison* of two variables between two groups, as might be seen in comparing treatment A, classically the experimental group, with treatment B, the control group of standard treatment or placebo.[3] For example, the independent variable (treatment A) results in the dependent variable (outcome in group A), and the independent variable (treatment B) results in the dependent variable (outcome in group B). The comparison between the two outcome variables helps determine whether treatment A is better than treatment B.

The specific focus of the present tutorial (*Part A*) is to explain the concepts of the null hypothesis and statistical significance as they apply to statistical indexes of contrast (Student *t*, Mann-Whitney *U*, and χ^2 tests) that are used in comparisons between two groups. The second tutorial (*Part B*) will address the application of these three common statistical tests.

At the end of each subsection, a *summary* is provided to assist the reader in a rapid overview of the report. Formulas and calculations generally have been omitted from the text but are provided in Tables I–V and in footnotes for interested readers. Figures 1–6 attempt to characterize the important basic concepts in easy-to-understand, easy-to-remember graphic illustrations.

TABLE I.
Variability *Within* a Single Group.[3, 5]

	Sum of Squares (SS)*	Variance (s^2)	Standard Deviation (s)	Standard Error (SE)
Continuous	$\sum (X_i - \bar{X})^2$	$\sum (X_i - \bar{X})^2/N - 1$	$\sqrt{\sum(Xi - \bar{X})^{(2)}/n - 1}$	$s/\sqrt{N} = \sqrt{(s^2/N)}$
Proportions	npq	npq/n = pq	√pq	√(pq/n)

*The sum of squares of a *dependent variable* may be also known as the total sum of squares (TSS), which is the total variation that a statistical model must attempt to explain.[2]

Σ, sum; X_i, individual value; $\bar{X}$, mean of the group; N, number of observations in group; SE, standard error of the central index (mean or proportion); p, successes in a group; q, 1 − p (failures in a group).

Fundamental Concept of Null Hypothesis

Most statistical tests are based on a fundamental concept of rejection or conceding the *null hypothesis* (H_0), the hypothesis that no difference between test groups really exists. In research, only *samples* can be taken. The complete *population* of the universe cannot be tested. Even though the contrast between two treatment regimens might suggest that treatment A is better than treatment B, it must be remembered that this suggestion is the result of contrasting two samples, sample group A versus sample group B, not the result of contrasting two universal complete populations; the apparent difference may be due to random variation (chance).

As we deal with samples and attempt to generalize to whole populations, we must decide at what point we will *reject the null hypothesis* (H_0) and concede the possibility of an *alternate hypothesis* (H_1), the hypothesis that a difference really exists. If we observe from our samples that treatment A seems to be quite different from treatment B, we are tempted to reject the null hypothesis and conclude that there really is a difference between these two groups.

Because we plan to test the null hypothesis using only samples, we are forced to decide at what point the sample data are sufficiently different to be *inconsistent* with the null hypothesis and allow us to reject it. We must choose a level of risk of error with which we can live. In clinical research, that point is often 5%. That means we are willing to accept the risk of a *type I, or false-positive, error* 5% of the time if we reject the null hypothesis. This point is called an *alpha level* (α), and the notation is generally that statistical significance will be recognized if $P \leq .05$; this means that if the probability (P) of being wrong is equal to or less that 5%, we will reject the null hypothesis.

In summary, the null hypothesis postulates that there is no difference between the samples being compared because both samples come from the same parent population.

General Concepts of Statistical Significance

It is easier to explain statistics beginning with continuous data and moving to ordinal and dichotomous data; therefore, we use this approach.

Student* t *test for continuous data. When comparing two groups on a continuous variable, the determination of statistical significance depends on 1) the magnitude of the observed difference and 2) the amount of *spread, or variability, of the data*, characterized as a frequency distribution around a *central index*, such as the mean (Fig. 1). If the data are tightly distributed about the central index and the distributions of the two groups of outcomes do not overlap, the two groups are obviously different (Fig. 2A). However, this scenario rarely occurs.

The central indexes (eg, the means or medians) of two groups are generally different; however, the bodies of the data in the two groups usually overlap (Fig. 2B). The question is, just how large is this difference between means compared with the spread, or variability, of the data? In other words, how does the difference *between* the groups compare with the variability *within* each group?[2,4] Table I lists common formulas for defining the variability *within* each group.

Because the spread, or variation, of the data is so important for determining statistical significance, a method was developed to standardize values free of the original measurement units. This effort generated the concept of the *Z-score* to describe single-sample data, where Z_i = ($Zi = (Xi - \bar{X})/s.$). This describes how many

TABLE II.
Parameters of Groups.

Treatment Group	Mean	Standard Deviation*
A	$\bar{X}_A$	s_A
B	$\bar{X}_B$	s_B
Difference between the means	$\bar{X}_A - \bar{X}_B = \bar{X}_C$	√[(sA2/NA) + (SB2/NB)] = SED

*The standard deviation of a group of difference between means is, in fact, the standard error of the difference.

($\bar{X}_A - \bar{X}_B$, difference between the means of two sample groups; $\bar{X}_C$, mean of the new hypothetical probability distribution curve of the differences between means; N_A, number of observations in group A; N_B, number of observations in group B; s_A^2, variance of group A; s_B^2, variance of group B; SED, standard error of the difference between the means.

TABLE III.
Calculation of Z-Scores and Z- and *t*-Test Values.

Single sample group data assumed to be a normal frequency distribution and probability curve (calculating position for each value)	$Z_i = (X_i - \bar{X})/s$
Null hypothesis assumption frequency distribution and probability curve of difference between two sample means (calculating Z- or *t*-test values)	Z or $t = (\bar{X}_A - \bar{X}_B)/SED$

Z_i, Z-score of an individual observed value; X_i, individual observed value; $\bar{X}$, mean of the single sample group data; s, standard deviation of the single sample group data; $(\bar{X}_A - \bar{X}_B)$, difference between the means of two sample groups; SED, standard error of the difference between the means.

standard deviations (s) units (SD units) from the mean $((\bar{X}))$ is the observed value (X_i). (Standard deviation is explained in detail in the previous tutorial [Part VI[1]]). The Z-score is the number of *SD units* from the mean. For example, the Z-score of a value exactly at the mean is 0. If the observed value is 1 SD unit from the mean, then the Z-score is 1.00; if the observed value is 2 SD units from the mean, the Z-score is 2.00[2,3,5] (Fig. 3).

The Z-scores assume that the frequency distribution of the data inscribes a normal gaussian or "bell" curve. The area under the curve demarcated by Z-scores may be calculated to give the proportion of the distribution included within these scores. For example, in a gaussian distribution, the zone demarcated by one Z-score above and below the mean includes 68.3% of the data; two Z-scores above and below the mean demarcate 95.4% of the data; and three Z-scores above and below the mean demarcate 99.7% of the data. To include exactly 95% of the data, 1.96 Z-scores above and below the mean are required.

Z-score values may be used to demarcate the boundaries between the "*inner zone*" and the "*outer zone*" (Fig. 3). An inner zone implies that all of the data in it *characterizes the entire group*. For example, if we sought a "*range of normal*" for blood pressure, we might consider taking the central 95% of all the blood pressures we have observed in a large general sample.[3] Any blood pressure values that fall outside the inner zone into the outer zone might be considered abnormal and therefore might be considered not to belong to the normal group.

The nice thing about normal distributions and Z-scores is that they also give information about probability.[3] Using the preceding example, if we obtained a blood pressure value from someone who appeared to be healthy, we might expect a 95% chance that the value would fall within the range of normal and would have a 5% chance of falling outside that range. The next few paragraphs are important for understanding the *leap from describing one group of sample data to comparing two samples* to determine the difference between them.

The Z-score concept, which was designed to demarcate a single-sample distribution, may be used when two samples are compared, such as treatment A versus treatment B. When the means of the two groups are subtracted $((\bar{X}_A - \bar{X}_B = \bar{X}_C))$, the result is the mean of a *new group* describing the *difference between the sample means*, with the *new group mean* of $\bar{X}_C$ and with a *new standard deviation*, known as the *standard error of the difference (SED)*[3] (Figs. 4 and 5 and Tables II and Table III).

Because the precomputer era required faster and more efficient methods than laborious by-hand calculations of multiple direct differences, the process of determining statistical significance classically became inferential, meaning significance was inferred from a null hypothesis and the assumption that the data belonged to some theoretical distribution, usually the normal distribution described by Gauss (ie, a gaussian distribution).[2,3,5] This means that if multiple samples from the parent population A and from the parent population B were hypothetically taken, the multiple differences of the many pairs of means of these samples would generate a

TABLE IV.
Standard Error of the Difference (SED).

	Standard Error of Differences
Continuous (differences between means for Z or *t*-tests of means)	$SED = \sqrt{[s_A^2/N_A) + (s_B^2/N_B)]}$
Proportions (differences between proportions)	$SED = \sqrt{[[\bar{p}(1 - \bar{p})][(1/N_A) + (1/N_B)]}$
	$SED = \sqrt{[NPQ/n_A n_B]}$

Classically, the standard error is standard deviation (s) divided by the square root of N; it is also the square root of the dividend of the variance (s^2) divided by N. In this situation, in which the variances are combined in this special way, the standard error of the difference is the square root of the sum of the two variances, each divided by the number of observations in the respective groups.

s_A^2, variance of group A; s_B^2, variance of group B; N_A and n_A, number of observations in group A; N_B and n_B, number of observations in group B; $\bar{p}$, all "successes" in both groups/total number of all observations in both groups; "Variance of a difference = sum of the individual variances",[2] variance of A group plus variance of B group.

TABLE V.
2 × 2 Contingency Table.

	Outcome		
Interventions	Success (+)	Failure (−)	
Group A	a	b	Total row (n_A) a + b
Group B	c	d	Total row (n_B) c + d
	Total column (f_1) a + c	Total column (f_2) b + d	Total (N) a + b + c + d

There are two rows, labeled Group A and Group B.
There are two columns, labeled Success (+) and Failure (−).
Internal cells, which contain observed data, are a, b, c, and d.
The marginal cells (shaded) (f_1, f_2, n_A, n_B, and N), are the totals of the column and row values, respectively.

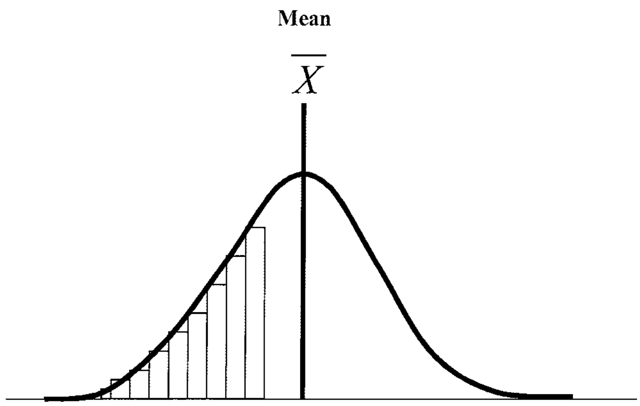

Fig. 1. Graphic illustration of a frequency distribution curve of continuous data from a single sample. The Y-axis indicates the frequency of occurrence of a specific value or a bin of values, and the X-axis indicates the specific values or bins (or aggregates) of values in the distribution.

normal *distribution curve of differences between means*, with the estimate of the mean of this new group to be $\bar{X}_C$.[3] Because the distribution is one of means, rather than individual observations, the standard deviation of this distribution is, in fact, a standard error, in this case the standard error of the difference (SED) between means.[2] This new group will have a normal, gaussian frequency distribution, which is known as a *parametric* distribution, meaning that the curve may be fully constructed from its *parameters*, the mean and standard deviation (in this case, the SED).[3]

The new group distribution curve may be constructed using the new parameters and ($\bar{X}_C$ and SED) and subdivided into standard error (SE) units.* Percentages of the distribution area under the curve may be defined by using the previously explained Z-score concept. However, in this case, rather than the SD units, SE units are used. To include exactly 95% of the probable range of the real mean difference, 1.96 SE units above and below the mean are required[2] (Figs. 4 and 5).

Just as the "range of normal" might be defined as the central 95% of the frequency distribution of data from one sample group, the central 95% of the frequency distribution of differences between the means might be considered as the "range of normal" for the *null hypothesis*. This means that any observed difference between the means that falls within this inner zone is considered a member of the null hypothesis (not statistically significantly different from the null hypothesis) and anything outside this zone is suspected of not belonging to the null hypothesis (ie, it is statistically significantly different from the null hypothesis[2–5]) (Figs. 4 and 5).

Because the critical issue is what values fall outside the inner zone, the *focus in statistical significance is the outer zone*. Thus, the boundary of this point between the inner zone and the outer zone (or zones) depends on *alpha (α)*; this boundary is identified as *Z-alpha* or $Z\alpha$. If alpha is set at 0.05, Z-alpha ($Z\alpha$ or $Z_{0.05}$) is ± 1.96 for a *two-tailed arrangement* to include the central 95% of the data. For a *one-tailed arrangement* with alpha set at 0.05, Z-alpha is ± 1.645 to demarcate 95% of the probable values of the difference between means on the larger side and 5% on the smaller side. In Figure 5, only a one-tailed test on the positive side (right side) of the distribution curve is shown; the test could be performed on the negative side (left side) of the distribution curve. Because the results of a comparison can usually go either way, a *two-tailed arrangement* is the used for most clinical research.

Alpha (α) is set before the study is begun and represents the proportion of the hypothetical distribution of the null hypothesis that we want to consider as "outsiders" and as not part of the "range of normal." We have an α chance of being wrong if we call the difference between the two groups (treatment A versus treatment B) an "outsider" and reject the null hypothesis. We could have chosen alpha (α) to be something else (eg, $\alpha = 0.01$, to demarcate 99% as $1 - \alpha$ and 1% as α). Why not just set all observations with a small alpha (eg, $\alpha = 0.01$)? Because the smaller the alpha, the larger the risk of a beta (false-negative) error and the larger the sample size required.

Mann-Whitney* U *test for ordinal data. Ordinal, nominal, and dichotomous data are generally managed in a *nonparametric* manner. Ordinal, nominal, and dichotomous data and small samples may not be analyzed using inferential parametric methods as described earlier in the present tutorial. When one cannot make assumptions about the relationship of the observed sample data and the hypothetical parent general population, the data must be analyzed without reference to a theoretical gaussian or other distribution. In nonparametric procedures the null hypothesis distributions are determined directly by permutations of the actual sample values or by permutations of the *ranks* of those values.[3]

The *Mann-Whitney* U *test* is a powerful nonparametric test used for ordinal data by comparing the *ranks* of the values in each group. (Thus, it is conceptually similar to a Student *t* test for ranked data). If the ranks of the outcomes in treatment group A were all greater than the ranks of the outcomes in treatment group B, it would be clear that the treatments gave significantly different results. On the other hand, if the ranks of subjects' outcomes overlapped, it would be difficult by merely looking at the results to determine whether or not the groups differed significantly.

The Mann-Whitney *U* test starts by ranking all the *values* together from the lowest to the highest; if ties are found, it takes the average value of the ranks included in

* As illustrated in Figure 6, showing two possible results of the comparison between two groups, when the difference between the means is the same in the two examples but spread of the data is larger in one (the SED is larger), the more the $((\bar{X}_A - \bar{X}_B)/SED)$ result moves to the center; conversely, the smaller the SED, the more the result moves to the tail (Fig. 6). This $((\bar{X}_A - \bar{X}_B)/SED)$ result is known as the *critical ratio*, which explores the relationship of the *difference* between *groups* to the *variation* within *groups*. It is called a *critical* ratio because it is used to make the decision of whether or not to reject the null hypothesis. The general formula for a critical ratio is as follows: *Parameter / SE of that parameter*. For continuous data analyzed by the Student *t* test or z-test for means, the *t* or *z* statistic produced is the critical ratio, which is the difference between the two means/the SE of the difference between the two means. The same principle can be applied to proportions (Table IV).

† $\chi^2 = \Sigma$[(observed − expected)^2/expected] This means χ^2 is the sum of this calculation [(observed − expected)^2/expected] for each cell in a contingency table.

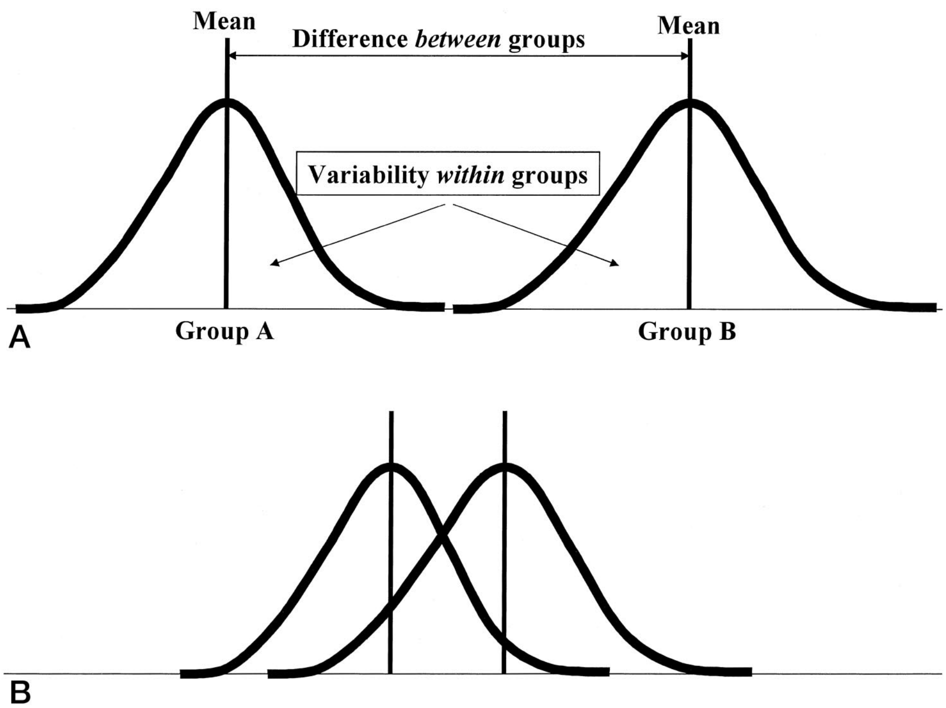

Fig. 2. (**A**) Graphic illustration of two treatment groups which are so completely different that there is no overlap. The degree of the difference is seen by noting that the difference *between* the groups' means is great and that the variability, or spread, of the data *within* the groups does not overlap. (**B**) Graphic illustration of the more common situation in which the means are different; however, the spread of the data *within* two treatment groups overlap. Statistical tests of the null hypothesis are required to determine whether the two treatments are truly different or just appear different because of chance.

the ties and gives that average value to each of the ties. It then separates these ranks into the original treatment groups and sums the ranks. One group sum is R_1, and the other is R_2. For mathematical reasons, which are not explained in this tutorial, the next step is to calculate U_1 and U_2 as follows: $U_1 = R_1 - [n_1(n_1 + 1)/2]$ and $U_2 = R_2 - [n_2(n_2 + 1)/2]$, wherein n_1 and n_2 represent the number in each group (group 1 or 2 [or group A or B]), respectively. The *smaller* of the two U values is taken as the test statistic U.[4,5]

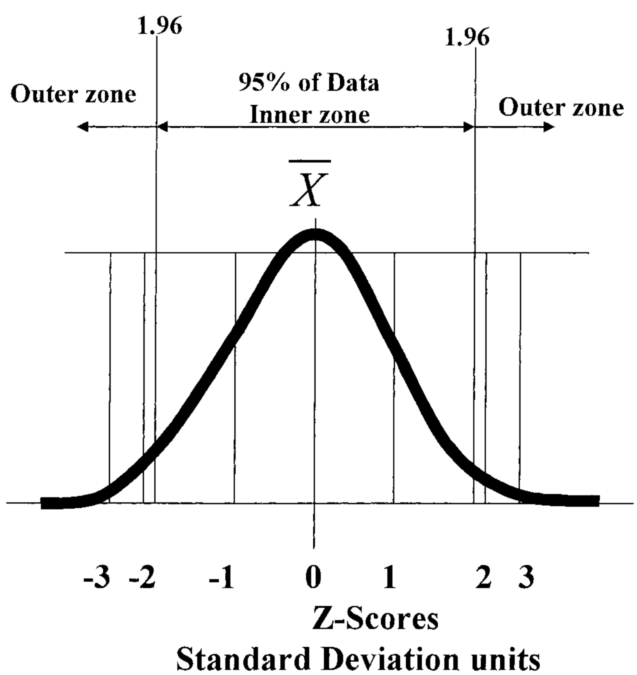

Fig. 3. Graphic illustration of a single-sample frequency distribution curve showing demarcations by Z-scores into an inner zone (comprising 95% of the area under the curve) and two outer zones (comprising 2.5% of the area under the curve in each tail).

The U statistic is compared with a table of U values that are associated with P values; this is performed either by hand or by a computer statistical software program. As mentioned in the preceding section on continuous data, the P value determines whether the difference between these samples is in the "range of normal" for the null hypothesis or is far out into the tail and thus probably not a good representative of the null hypothesis. In the latter case, we would reject the null hypothesis with a P probability of being wrong. Notice, once again, that the P value does two things: 1) It locates our sample comparison on a probability distribution curve and 2) it tells us just how wrong we would be if we rejected the null hypothesis.

Chi-square test for nominal and dichotomous data. Nominal and dichotomous data are generally organized into frequency counts, which generate proportions. The chi-square (χ^2) test is a popular way to compare proportions between two (or more) groups. The χ^2 statistic derives from the comparison between what might be *"expected" from the null hypothesis* and what is actually observed.† Once calculated from the data, the χ^2 statistic is compared with a table of χ^2 distributions that are associated with P values. The P value determines whether the difference between the two treatments is indistinguishable from the null hypothesis or whether it is so far out into the tail of the distribution as to be inconsistent with the null hypothesis. In such a case, we would reject the null hypothesis with a P probability of being wrong.[2,5]

The χ^2 test is easier to understand by looking at a 2 × 2 contingency table (Table V). The null hypothesis

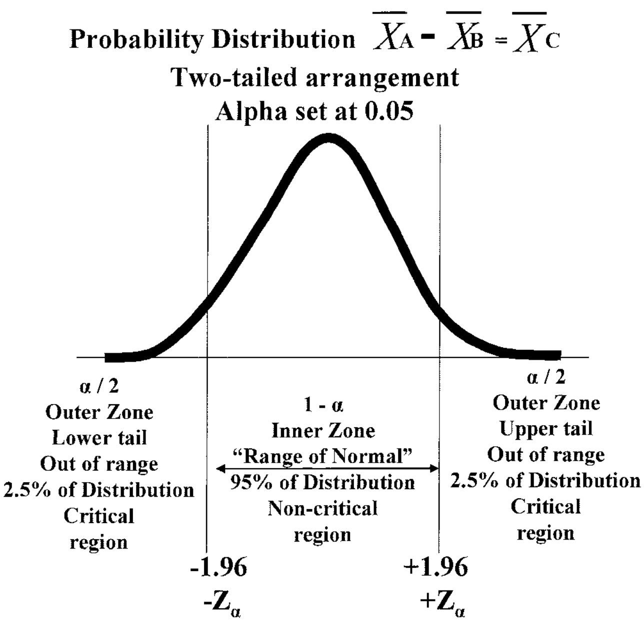

Fig. 4. Graphic illustration of a probability curve of the comparison between two groups. The new mean of this curve is the difference between the means of group A and group B. The variability, or spread, of the curve is the pooled variance of the two samples. The curve is demarcated by standard errors of the difference between the two samples and is shown in a two-tailed arrangement, in which the central 95% of the difference between the means is bordered on both sides by two critical regions of differences inconsistent with the null hypothesis. The two-tailed arrangement is the usual one because which of the two treatments might be better cannot be predicted until tested.

states that both samples (those in treatment group A and those in treatment group B) are from the same parent population and that differences between them are only due to chance and are not due to the treatments. The marginal cells (shaded in Table V) are estimates of this parent population; this means that all the successes (f_1) and all the failures (f_2) as a proportion of the total of both samples (N) represent the proportion of successes (f_1/N) and failures (f_2/N) under the null hypothesis. Thus, one would expect to see this same proportion of successes and failures in each treatment group relative to the number of subjects in that group. For example, the "expected" number of successes in cell "a" would be $f_1/N \times n_A$ or, configured another way, $f_1 n_A/N$. Multiplying the margin value of the column by the margin value of the row and dividing by the total N generates the expected values of each cell. These are the *"expected"* number of observations in each cell, but they can contain fractions of a number and must remain fractions. The expected values also must add up to the appropriate row and column totals.

The *actual observed values* of numbers of successes and failures in each treatment arm are the internal cell values (a, b, c, and d). With this information, the χ^2 value can be calculated as seen in the preceding footnote.[5]

In summary, the "range of normal" (usually the central 95% of the frequency distribution) or the "expected values" characterize the majority of the null hypothesis population, which means that both samples being compared come from the same parent population and that, therefore, there is no real difference between them. Statistical significance is said to occur when the difference between the compared samples is so far out in the tail as to be an "outsider" and thus *inconsistent with the null hypothesis*.

CONCLUSION

Comparative research is based on the comparison of *samples*, rather than the comparison of universal *populations*. Therefore, any study, such as treatment A versus treatment B, attempts to *generalize* from samples (group A, group B) to populations. This means that if the samples show a significant difference in the direction of sample A, then a similar difference would be expected between many more samples, and even between the universal populations; therefore, with high confidence we can be fairly sure that treatment A is the better treatment.

It is generally safer to assume that there is no difference between the populations and that chance alone would cause samples to be different; this is what is meant by the *null hypothesis*. The process of comparative statistics tests the null hypothesis with the sample data. If the sample data shows a large enough *difference* between the two samples as to be *inconsistent* with the majority of the

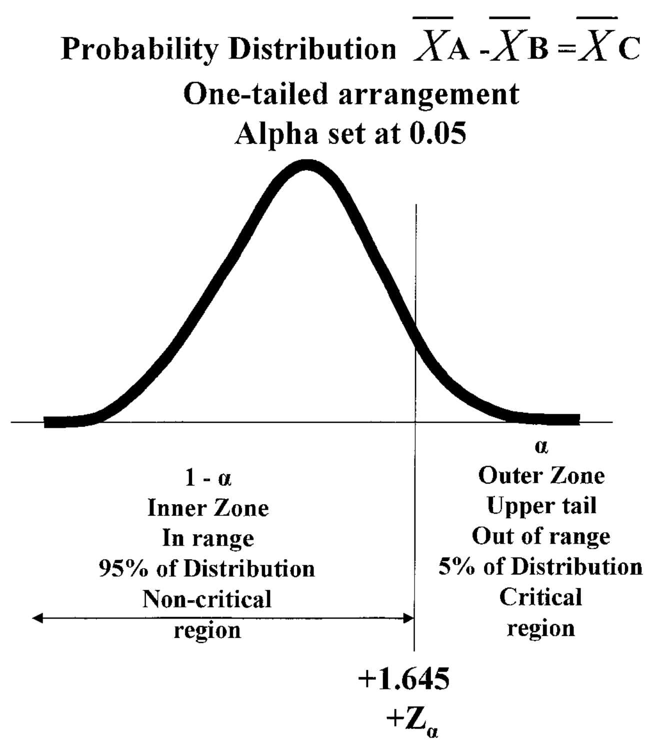

Fig. 5. Graphic illustration of a probability curve of the comparison between two groups. The new mean of this curve is the difference between the means of group A and group B. The variability, or spread, of the curve is the pooled variance of the two samples. The curve is demarcated by standard errors of the difference between the two samples and is shown in a one-tailed arrangement, in which 95% of the difference between the means is on one side of the curve and the critical region of the differences inconsistent with the null hypothesis is in one tail. This arrangement is used for rare conditions in which the results can only sensibly go in one direction.

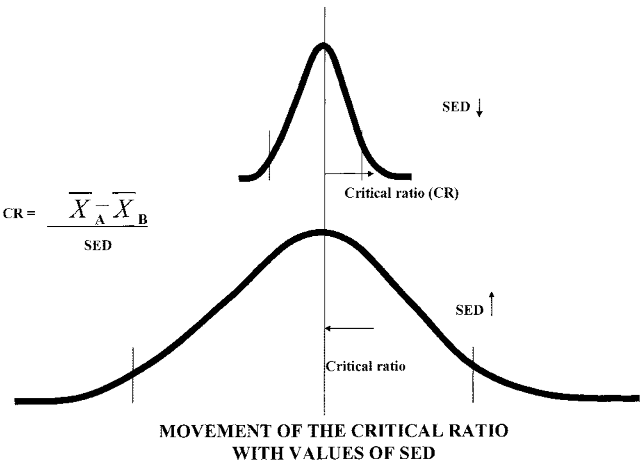

Fig. 6. Graphic illustration attempting to show how important the spread of the data is. Knowing the difference between the means is only one part of the equation that determines whether the difference between the two samples is statistically significant. Given a specific difference between the means, if the standard error of the difference (SED) is small, as in the top curve, there is a good chance that the critical ratio will be out in the tail and thus statistically significant. On the other hand, as illustrated in the bottom curve, if the SED is large, it is probable that the critical ratio will fall within the majority of the values consistent with the null hypothesis.

values that fit well within the null hypothesis "range of normal" (often the central 95% of the standard errors about the mean of the difference), we can reject the null hypothesis for that specific experiment and conclude that there really is a difference between treatment A and treatment B that is not just due to chance.

The tests of statistical significance for comparisons determine how far out toward the tail of the distribution curve of the differences under the null hypothesis assumption the observed sample difference fits. If the sample difference is far out in the tail, it would seem to be an outsider to the majority of values of the null hypothesis and thus, statistically significantly different from the null hypothesis.

The primary measure of the statistical difference is known as the critical ratio, which is the difference between the central indexes of the two samples divided by the SE of the difference; this number is the *statistic*. The *statistic* is associated with a probability (P), which indicates how probable is it that we would be wrong if we rejected the null hypothesis. Prior to beginning the study, an alpha level (á) is set, often at a level of 0.05; this level indicates the point at which we would be comfortable in making an error if we rejected the null hypothesis. Thus, if the P value associated with the *statistic* calculated from the sample differences is less than the preset alpha, statistical significance has been achieved; and by the preset choice, we would reject the null hypothesis and conclude that the sample difference is really the result of the treatments and not due to chance.

Acknowledgments

The authors thank Dr. Neely's long-time friend and mentor, James F. Jekel, MD, MPH, Professor of Epidemiology and Public Health, Emeritus, Yale University School of Medicine, who was kind enough to take time from his "retirement" to critically review and participate in revising these tutorials.

BIBLIOGRAPHY

1. Neely JG, Stewart MG, Hartman JM, Forsen JW, Jr., Wallace MS. Tutorials in clinical research, part VI: descriptive statistics. *Laryngoscope* 2002;112:1249–1255.
2. Jekel JF, Katz DL, Elmore JG. *Epidemiology, Biostatistics, and Preventive Medicine*, 2nd ed. Philadelphia: WB Saunders; 2001:137–208.
3. Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia: WB Saunders; 1985: 89–169.
4. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2000:387–426, 473–490, 537–556.
5. Feinstein AR. *Principles of Medical Statistics*, 1st ed. New York: Chapman and Hall/CRC; 2002:23–346, 489–514.