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

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. The present tutorial is the first of two tutorials concerned with bivariate analysis, a process in which the relationship between two variables is explored. 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).

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. 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 $\chi^2$ 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.
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 ($\alpha$), 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? 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 $Z_i = (\bar{X} - \bar{X})/s_\bar{X}$. This describes how many

---

**Table I. Variability Within a Single Group.$^3, 5$**

<table>
<thead>
<tr>
<th>Continuous</th>
<th>Sum of Squares (SS)$^*$</th>
<th>Variance ($s^2$)</th>
<th>Standard Deviation ($s$)</th>
<th>Standard Error (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportions</td>
<td>$\sum (X_i - \bar{X})^2$</td>
<td>$\sum (X_i - \bar{X})^2/N - 1$</td>
<td>$\sqrt{\sum (X_i - \bar{X})^2/N - 1}$</td>
<td>$s/\sqrt{N} = \sqrt{(s^2/N)}$</td>
</tr>
</tbody>
</table>

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.$^2$

$\sum$, 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).

---

**Table II. Parameters of Groups.**

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Mean</th>
<th>Standard Deviation$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\bar{X}_A$</td>
<td>$s_A$</td>
</tr>
<tr>
<td>B</td>
<td>$\bar{X}_B$</td>
<td>$s_B$</td>
</tr>
<tr>
<td>Difference between the means</td>
<td>$\bar{X}_A - \bar{X}_B = \bar{X}_C$</td>
<td>$\sqrt{(s^2A/NA) + (s^2B/NB)} = SED$</td>
</tr>
</tbody>
</table>

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; $s^2_A$, variance of group A; $s^2_B$, variance of group B; SED, standard error of the difference between the means.
The nice thing about normal distributions and Z-scores is that they also give information about probability. 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 ((\(X_A - X_B\)), the result is the mean of a new group describing the difference between the sample means, with the new group mean of \(X_C\) 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 III.**

<table>
<thead>
<tr>
<th>Calculation of Z-Scores and Z- and t-Test Values.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single sample group data assumed to be a normal frequency distribution and probability curve (calculating position for each value)</td>
</tr>
<tr>
<td>(Z = (X_i - \bar{X})/s)</td>
</tr>
<tr>
<td>Null hypothesis assumption frequency distribution and probability curve of difference between two sample means (calculating Z- or t-test values)</td>
</tr>
<tr>
<td>(Z \text{ or } t = (\bar{X}_A - \bar{X}_B)/\text{SED})</td>
</tr>
</tbody>
</table>

---

**TABLE IV.**

<table>
<thead>
<tr>
<th>Standard Error of the Difference (SED).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous (differences between means for Z or t-tests of means)</td>
</tr>
<tr>
<td>(\text{SED} = \sqrt{s_A^2/N_A + (s_B^2)/N_B} )</td>
</tr>
<tr>
<td>Proportions (differences between proportions)</td>
</tr>
<tr>
<td>(\text{SED} = \sqrt{p(1-p)(1/N_A + 1/N_B)} )</td>
</tr>
<tr>
<td>(\text{SED} = \sqrt{NPQ/n_A n_B} )</td>
</tr>
</tbody>
</table>

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 (\(\sigma^2\)) 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; \(p\), 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.**

<table>
<thead>
<tr>
<th>2 x 2 Contingency Table.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interventions</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Group A</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Group B</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total column (f1)</td>
</tr>
<tr>
<td>a + c</td>
</tr>
</tbody>
</table>

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) \(f_1, f_2, n_A, n_B, \text{ and } N\), are the totals of the column and row values, respectively.

---

Laryngoscope 113: September 2003

Neely et al. Comparative Statistics

1536
normal distribution curve of differences between means, with the estimate of the mean of this new group to be $X_C$. 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. 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).

The new group distribution curve may be constructed using the new parameters and $(X_C$ 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 (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) (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 ($\alpha$); this boundary is identified as $Z$-alpha or $Z$. If alpha is set at 0.05, $Z$-alpha ($Z_{0.05}$) is $\pm 1.96$. 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 $\pm 1.645$ 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 ($\alpha$) 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 $\alpha$ 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 ($\alpha$) to be something else (eg, $\alpha = 0.01$, to demarcate 99% as $1 - \alpha$ and 1% as $\alpha$). 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 parental 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.

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 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...
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:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)/2}{n_1}$$

and

$$U_2 = R_2 - \frac{n_2(n_2 + 1)/2}{n_2}$$

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 $U$.

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 ($\chi^2$) test is a popular way to compare proportions between two (or more) groups. The $\chi^2$ 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 $\chi^2$ statistic is compared with a table of $\chi^2$ 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.

The $\chi^2$ test is easier to understand by looking at a $2 \times 2$ contingency table (Table V). The null hypothesis
states that both samples (those in treatment group A and those in treatment group B) are 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 samples, the hypothesis is rejected.
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 falls. If the sample difference is far out in the tail, it would seem to be an outlier 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 (α) 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