0363-5465/97/2525-0882$02.00/0
THE AMERICAN JOURNAL OF SPORTS MEDICINE, Vol. 25, No. 6

Current Concepts

A Statistics Primer

Tests for Continuous Data

Mary Lou V. H. Greenfield,* MPH, MS, Edward M. Wojtys, MD, and John E. Kuhn, MS, MD

From MedSport and the Section of Orthopaedic Surgery, University of Michigan, Ann Arbor, Michigan

When we read research studies or when we design experiments, a common question frequently arises: Are the populations being studied and compared with one another similar? If any differences are present, are these differences a result of the treatment effect or study intervention? Because studying the entire population is generally not realistic, sample groups are drawn from the population of interest and the results are then generalized to the rest of the population. Statistical tests are available to help decide if the difference between either the means or proportions of sample groups (experiment and control) are statistically significant. As discussed in a previous "Current Concepts" article in this journal, if the outcome of interest includes data that fall into discrete categories (e.g., presence or absence of disease, gender, race), statistical tests for discrete data are used.[2] If the outcome of interest is a comparison of sample means for data that are continuous (e.g., height or weight of two populations) then statistical tests for continuous data are used.[1] Specific tests such as the Student's *t*-test and analysis of variance (ANOVA) are commonly used statistical methods for continuous data.

ONE-SAMPLE *t*-TEST

Consider the example of evaluating the weight of high school football players. Suppose a coach wanted to know if the mean weight of the players on his team is similar to the reported mean weight for all high school football players in the country. The coach would weigh all of the players on his team and obtain the mean weight (and standard deviation) of the team; he would then compare that weight with the known mean weight of players across the country. Because weight falls on a continuous numeric scale, the coach should investigate whether there is a bell-shaped (Normal) distribution of weights on his team. If the distribution of the data is symmetric (Normal), then he might compare the mean weight of his team with the countrywide mean weight using a one-sample *t*-test.

A *one-sample t-test* is used when the investigator wants to compare a sample mean value of some variable with a known value of some *standard* variable. This is accomplished by comparing the sample mean to the known value in the population relative to the standard error. (The mean and standard error are measures of central tendency and spread, respectively. A basic review of these concepts can be found in a previous "Current Concepts" article.[1]) In this example of a one-sample *t*-test, the coach compares his team's mean weight with the countrywide mean weight and then computes a *t-statistic*. (Formulas for computing *t*-statistics are beyond the scope of this article, but they can be found in any basic statistics text.) The *t*-statistic is the *critical value* of the *t*-test and is associated with a probability or *P* value. (The *P* value associated with the critical *t*-statistic can be found in the back of any statistics book.) Generally, if this *P* value is less than 0.05, the investigator has enough evidence to reject the null hypothesis that his team's mean weight is the same as the national mean weight of football players. In this example, if the *P* value associated with the one-sample *t*-test is 0.02, this would indicate a statistically significant difference between the coach's high school players' weights and the countrywide weight. The actual mean weight must be noted to determine if the average member of his team is lighter or heavier than the average player countrywide.

INDEPENDENT TWO-SAMPLE *t*-TEST

In another example, suppose that the coach suspects that the archrival team across town has much heavier players

* Address correspondence and reprint requests to Mary Lou V. H. Greenfield, MPH, MS, University of Michigan, Orthopaedic Surgery, TC2914G-0328, 1500 East Medical Center Drive, Ann Arbor, MI 48109.
No author or related institution has received any financial benefit from research in this study.

than those on his team. The coach addresses this concern by studying the mean weight of players from both teams. This is a setting in which an *independent two-sample t-test* can be used. The two-sample t-test compares the mean values between two independent groups. (The groups are independent because none of the players on the coach's team are also on the rival team.) Similar to the one-sample t-test, the assumptions are that both teams have the Normal distribution of weights and equal standard deviations. The test is computed using a pooled standard error. The two-sample t-test also has a critical value associated with a probability or P value. Again, if this P value is less than 0.05, the coach has enough evidence to show statistically that the weights are not the same. Suppose in this example that the mean weight of the players on the coach's team was 210 pounds and the mean weight of those on the archrival's team was 220 pounds, with a critical t-statistic associated with a P value of 0.01. This would indicate a statistically significant difference between the average weight of the players on both teams. The coach would conclude that his players weigh significantly less than his archrival's players.

PAIRED *t*-TEST

Suppose a coach wants to know how his players' weights are affected by two practices per day in the preseason. To evaluate this training, the coach weighs the members of his team at both the beginning and end of a 2-week training session to monitor the effects of vigorous training on weight. At first, the coach considers conducting a two-sample independent t-test; that is, he considers comparing the team's mean weight before and after the 2-week training session. This is not the correct test to use because, in this example, the two samples are *not independent*. That is, the *same* players are being measured before and after training. *When the samples are not independent, a paired t-test is required.* The *paired* t-test measures the *difference* in the weight of each player before and after training and then uses the *mean difference* of each player to find the critical t-statistic and associated P value.

If the training has no effect, the mean difference in weight before and after training would be zero. This is the null hypothesis. If the t-statistic is associated with a significant P value (usually ≤ 0.05), then the null hypothesis may be rejected in favor of a real difference in weight before and after the training. If an independent two-sample t-test had been conducted, the average weight before and after training would not reflect the paired difference of each individual team member and could possibly result in the coach erroneously concluding that the training had no effect on players' weights, when there may have been one.

ANALYSIS OF VARIANCE (ANOVA)

The ANOVA is an extension of the t-test. In fact, when the means between *two* independent groups are compared, the conclusions from either the t-test or ANOVA are identical; that is, either test may be used to determine statistical significance. However, ANOVA is used to determine a significant difference in mean values *when there are more than two independent sample groups*. Specifically, as its name suggests, analysis of variance incorporates two pieces of variation in the data to determine statistical significance. The first source of variation summarizes the variability from mean to mean from *between groups* of observations. The second source is the *within-group variance*. It represents the variation from individual to individual. It can be calculated by determining the total variation of all of the observations from the overall mean and then subtracting the previously calculated *between-group variance*. The ANOVA compares the calculated variance *between* the means of all of the groups in the sample with the calculated variance pooled from *within* each of the groups. The ratio between these two variances is called the *F-ratio* or *F-statistic*. We would expect this ratio to be close to 1.0 if there is no difference among the groups we are studying. The F-statistic has a specific probability distribution, which is used to calculate the P value. This distribution is influenced by degrees of freedom. For our purposes it is important to note that the F-test is an omnibus test; that is, if the F-statistic is significant ($P < 0.05$) then we know that at least one of the group means is different.

For example, suppose a coach wants to study the weights of team members by year in school—9th-, 10th-, 11th-, and 12th-grade athletes. The null hypothesis is that the mean weights for players will be equal across all grades; the research hypothesis is that at least one of the mean weights is not equal to the others. The weights of players from each of the four groups are measured and an ANOVA is computed for these data. The F-statistic is associated with an overall P value of 0.03. This indicates that at least one of the groups of players has a mean weight that is different from the other group means; therefore, the coach rejects the null hypothesis in favor of the research hypothesis.

However, there is still a problem in this example because the omnibus F-test does not tell the coach which group mean is different from the other means, or if all of the group means are different from the others, or anything in between. There are several *posterior tests* that are used to make specific comparisons between all of the possible pairs of means. For example, the coach can compare 9th with 10th graders, 9th with 11th graders, 9th with 12th graders, 10th with 11th graders, and so forth. These types of tests are called posterior because they are conducted after the null hypothesis has been rejected based on a significant P values associated with the omnibus F-test. Comparing two groups at a time is known as a *pairwise comparison*. The posterior tests should maintain the Type I error equal to 0.05 for the entire series of comparisons. This is important because the more comparisons that are made, the more likely the investigator is to find a statistically significant result due to chance alone. To reduce this risk, the investigator should make adjustments for multiple comparisons. Examples of specific tests that adjust for these multiple comparisons include the *Bonferroni* method, the *Student-Newman-Keuls* procedure, the *Tukey*

method, and the *Scheffe* method. It is important that the coach conducts posterior tests after rejecting the null hypothesis because these posterior tests allow him to make comparisons between all of the grades, and give associated *P* values that are adjusted for these multiple comparisons.

This example has been limited to what is referred to as *one-way design* or a *one-way analysis of variance;* this is because there is only one *factor* or characteristic— grade— that is being evaluated. Analyses that include more than one factor (for example, grade and position played) are also available and may be referred to as *factorial ANOVA* because more than one factor is being analyzed at the same time. Another type of ANOVA that may be seen is a *repeated measures ANOVA.* This type of ANOVA is commonly seen in health research literature. Repeated measures ANOVA is used when measurements of the variable of interest are taken more than once for each individual. For example, if the coach wanted to measure players' weights at several times throughout the season, a repeated measures ANOVA would be an appropriate test to use.

ACKNOWLEDGMENT

The authors thank Dr. M. Anthony Schork from the University of Michigan, School of Public Health, Department of Biostatistics, for his review of this article.

REFERENCES

1. Greenfield ML, Kuhn JE, Wojtys EM: Descriptive measures for continuous data. *Am J Sports Med 25:* 720–723, 1997
2. Kuhn JE, Greenfield ML, Wojtys EM: Statistical tests for discrete data. *Am J Sports Med 25:* 585–586, 1997