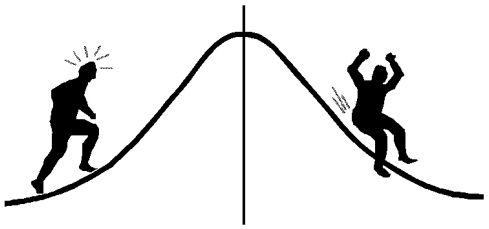

BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 21 DATE: August 18, 1989

SUBJECT: What are Degrees of Freedom?

Whenever you use a statistical test, degrees of freedom are necessary for determining the significance of the observed test statistic. But what are degrees of freedom? In this pamphlet I will discuss an operational definition.

Suppose you have twelve numbers, say the heights of twelve trees, and you ask someone to guess what those numbers are. If this is all you tell them then they could choose any twelve numbers. They would have twelve choices or twelve degrees of freedom. But suppose, when they quite naturally asked for more information, you gave them the mean. Now eleven numbers could be chosen in any way, but one of the numbers must be chosen so that the twelve numbers have the right mean. Since the numbers have been restricted to a specific mean, the set of numbers now has only eleven degrees of freedom, or eleven ways to vary. One degree of freedom has been devoted to the mean.

Generalizing from this, I define **degrees of freedom** as the number of possible values a set of numbers could have minus the number of restrictions placed on those numbers.

We place restrictions on data in order to simplify it. When we do this, we are **modeling**. Usually, we would rather discuss one number, the mean for example, than individual numbers. One number is usually easier to comprehend. But note that when we model the numbers by the mean, we are stating that the information contained in those numbers is essentially represented by that one number.

Following are some common statistical procedures that demonstrate this definition. Suppose that you also knew the ages of the twelve trees, and you were interested in a possible relationship between height and age. You might, for instance, postulate that there is a linear relationship between age and height so that if one tree is a year older than another it is some unknown amount (say, b) taller. A familiar model to fit is:

$$\text{Height} = \mathbf{a} + \mathbf{b}\ \text{Age}\ ,$$

where **a** is the intercept (a tree's height at age 0) and **b** is the slope (the difference in height between trees that are one year apart). Statisticians call these unknown values, **a** and **b**, **parameters** (the mean is also a parameter). With this model, we can represent the height data with just two numbers (the parameters), instead of twelve. The height data now has ten degrees of freedom. That is, ten of the heights can have any value, but the other two must be chosen so that all twelve numbers are best fit by the linear relation defined by **a** and **b**. Thus the residual mean squares has

ten degrees of freedom (or more generally, n-2 degrees of freedom, where n is the total number of observations in the dataset). The associated F-test has 1 and 10 degrees of freedom. The total degrees of freedom is still 12, since there is 1 for the mean (or **a**), 1 for **b** and 10 for the residual. The mean and intercept share a degree of freedom since **a** will be the mean if **b** = 0.

Continuing with the example of twelve tree heights, suppose that 4 trees had been given one kind of fertilizer, 4 had been given a second kind and the rest had been given no fertilizer (we will ignore the age information for now). Instead of representing or modeling the data with just one mean, three means might be more appropriate. Thus three of the twelve degrees of freedom are assigned to the means (the data have three restrictions placed on them). The nine remaining (residual df) indicate that nine of the heights can have any value.

We could also think of this as three sets of data, one for each treatment. Each set has one degree of freedom set aside for their mean. Suppose the sample sizes for each treatment were: 3,4, and 5 trees each. Then the error or residual degrees of freedom for each set is 2, 3, and 4, for a total of nine degrees of freedom. Regardless of the group sample sizes, the total residual degrees of freedom is nine. For example if the sample sizes had been 8, 1, and 3, then the residual degrees of freedom become 7, 0, and 2, for a total of 9.

The three means can be considered a set of numbers in their own right. They have three degrees of freedom. We can restrict them by requiring that their mean be equal to the grand mean (the original mean discussed above). The grand mean then takes one of their three degrees of freedom. Thus two means can have any value, but the third is restricted to whatever value is required to obtain the grand mean.

We can test if these three means are different from each other with an F-test (the usual analysis of variance). This test will have nine[1] degrees of freedom for the residual or denominator mean square and two[2] for the numerator mean square. Again, one degree of freedom is devoted to the grand mean.

The total of all the degrees of freedom in a statistical analysis will be equal to n, the number of observations. For those models or analyses we are most familiar with, one degree of freedom is always devoted to the grand mean. Then some degrees of freedom are associated with the parameters in the model which restrict the values the data can have. The leftover or residual degrees of freedom are associated with the variability in the rest of the data. This variability provides a measure with which to test the significance of the parameters of the model.

Suppose that we want to try a different model to our set of twelve numbers that includes age and the three treatments. We could speculate that each of the three groups should be fit with a

---

[1] or more generally, n-k, where k is the number of groups.

[2] or more generally, k-1.

simple linear regression of height against time. This model could use the following equations:

$$\text{Height} = \mathbf{a_1} + \mathbf{b_1}\ \text{Age} \quad \text{for group 1}$$
$$\text{Height} = \mathbf{a_2} + \mathbf{b_2}\ \text{Age} \quad \text{for group 2}$$
$$\text{Height} = \mathbf{a_3} + \mathbf{b_3}\ \text{Age} \quad \text{for group 3.}$$

This model has six parameters ($\mathbf{a_1, a_2, a_3, b_1, b_2, b_3}$) so that the degrees of freedom for the residual mean squares is six. If we restricted the slopes $\mathbf{b_1}$, $\mathbf{b_2}$, and $\mathbf{b_3}$ to have only one value instead of three (the usual analysis of covariance where the lines are assumed to be parallel), then the model would have four parameters ($\mathbf{a_1, a_2, a_3, b}$) and a residual degrees of freedom of eight.

Contingency tables provide another common type of statistic. Suppose that each of our twelve numbers was the number of trees assigned one of three treatments and classified as dead, poor, okay, and good at the end of the growing season. The familiar contingency table analysis restricts the treatment (or row) total counts to be the observed row counts, and the response categories (or column) counts to be the observed column counts. Also, the row counts and column counts themselves are restricted to add up to the total count. Thus there are r = 3 row counts minus 1 for the total count, plus c = 4 column counts minus 1 for the total count, and of course, 1 degree of freedom for the total count. The residual degrees of freedom (for the statistical test) is:

$$3 \times 4 - (3 - 1) - (4 - 1) - 1 = 12 - 2 - 3 - 1 = 6$$

or, in general terms:

$$r \times c - (r - 1) - (c - 1) - 1 = rc - r - c + 1 = (r-1)(c-1).$$

Thus there are 6 degrees of freedom for the χ^2-test.

CONTACT: Wendy Bergerud
387-5676

NEW PROBLEMS

What are the residual degrees of freedom for a simple regression on a dataset with 50 observations? How would the df change if you fit a multiple regression with three independent variables? How would the df for the multiple regression change if the dataset contained 3 numbers?

What are the residual df for a dataset with 70 observations which has been divided into 6 groups? What would be the df for the F-test of group differences?

What are the df for the usual contingency table χ^2-value for a dataset with 800 observations classified into a table with 3 rows and 6 columns?

What are the df for a t-test of a mean with a sample of 80?