

Michael S. Lewis-Beck
Department of Government, 224 Littauer
Harvard University
Cambridge, MA 02138
617-495-8280 (phone)
mlewis-beck@latte.harvard.edu
Word Count: 561

Degrees of Freedom. The number of independent observations available for PARAMETER ESTIMATION indicates the degrees of freedom, *df*. Generally speaking, a test statistic has degrees of freedom, determined by mathematical constraints on the quantities to be estimated. As examples, degrees of freedom must be calculated with the *T*-RATIO, the *F*-RATIO, the CHI-SQUARE, the STANDARD DEVIATION, or REGRESSION ANALYSIS. In the case of regression analysis, that number may be calculated by subtracting from the sample size the number of coefficients to be estimated, including the constant.

Consider the simple case of estimating the STANDARD DEVIATION of a population variable, from a random sample of N observations. The formula is as follows: $SD = \sqrt{\Sigma(U - \bar{U})^2 / N - 1}$, where SD = the standard deviation estimated in the sample, $\sqrt{\quad}$ = square root, Σ = sum, U = sample observations on the population variable, the \bar{U} = the average score of the sample observations on U , and N = the sample size. Note that the numerator is divided by $N - 1$, not N . Substantively, this adjustment may appear small. Nevertheless, it is theoretically important. The correction is necessary to avoid an exaggerated, BIASED estimate of the population standard deviation. In calculating this one population parameter, we have exhausted one degree of freedom. This is so because of the characteristics of the sample mean, used in the formula. There are N number of $(U - \bar{U})$ scores. However, just $N - 1$ of those are independent because once they are computed the last, the N th value, is necessarily fixed. This comes from the fact that

always $\Sigma (U - \bar{U}) = 0$. For instance, suppose $N = 3$, and $(U_1 - \bar{U}) = 6$, $(U_2 - \bar{U}) = 12$.

Then, it must be that $(U_3 - \bar{U}) = -18$. The restriction on the deviations from the mean, that they sum to zero, dictates that the last deviation is not “free” but fixed, so one degree of freedom is lost. In the example, only one population parameter, the standard deviation, is estimated, leaving degrees of freedom $= N - 1$.

In the more complicated case of regression analysis, the $df = N - K$, where $N =$ the sample size and $K =$ the number parameters to be estimated. Suppose the MULTIPLE REGRESSION model of $Y = a + bX + cZ + e$, where $N = 45$. There are three parameters to be estimated, the intercept a and the two slopes, b and c . Hence, $df = N - K = 45 - 3 = 42$. With regression analysis, it is important to have a sufficient number of degrees of freedom. Take the extreme, when the SIMPLE REGRESSION model $Y = a + bX + e$ is fitted to a sample with $N = 2$. Here the R -SQUARED $= 1.0$, because the straight line of the ORDINARY LEAST SQUARES fit connects the two data points without error. This is not perfect explanation but mere mathematical necessity, for the degrees of freedom have been exhausted, that is, $N - K = 2 - 2 = 0$. Generally, in regression analysis one wants many more independent observations than independent variables. When K begins to approach N , analysts pay special attention to the ADJUSTED R -SQUARED, which corrects for lost degrees of freedom.

MICHAEL S. LEWIS-BECK

References

Kennedy, Peter. (1998). *A Guide to Econometrics* (4th ed.). Cambridge, MA: MIT Press.

Lewis-Beck, Michael S. (1995). *Data Analysis: An Introduction*. Thousand Oaks, CA:

Sage.