Married Man: There is only one subject and my degree of freedom is zero. I should increase my "sample size."

Degrees of freedom

Dr. Alex Yu

"Degrees of freedom" have nothing to do with your life after you get married. Actually, "Degrees of freedom" (df) is an abstract and difficult statistical concept. Many elementary statistics textbook introduces this concept in terms of **the number that are "free to vary"** (Howell, 1992; Jaccard & Becker, 1990). Some statistics textbooks just give the df of various distributions (e.g. Moore & McCabe, 1989; Agresti & Finlay, 1986). Johnson (1992) simply said that degrees of freedom is **the "index number" for identifying which distribution is used**.

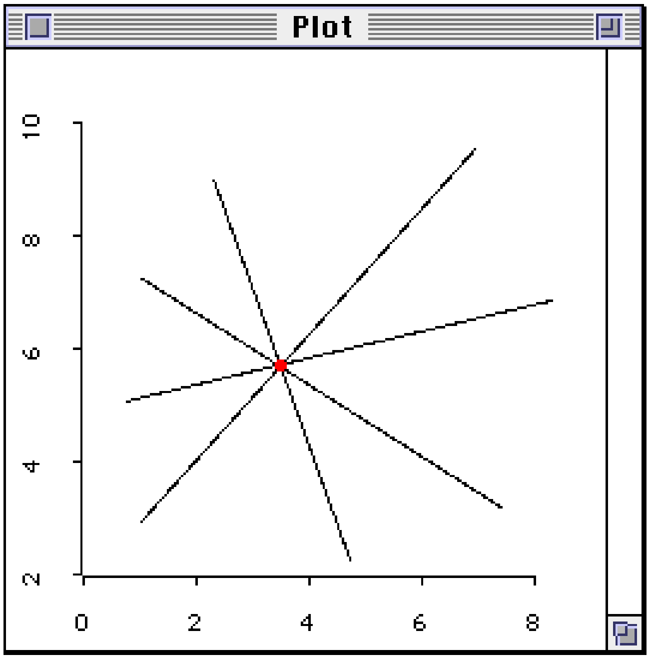
Probably the preceding explanations cannot clearly show the purpose of df. Even advanced statistics textbooks do not discuss degrees of freedom in detail (e.g. Hays, 1981; Maxwell and Delany, 1986; Winner, 1985). It is not uncommon that many advanced statistics students and experienced researchers have a vague idea of degrees of freedom. In this write-up I would like to introduce this concept from two angles: working definitions and mathematical definitions.

Working Definitions

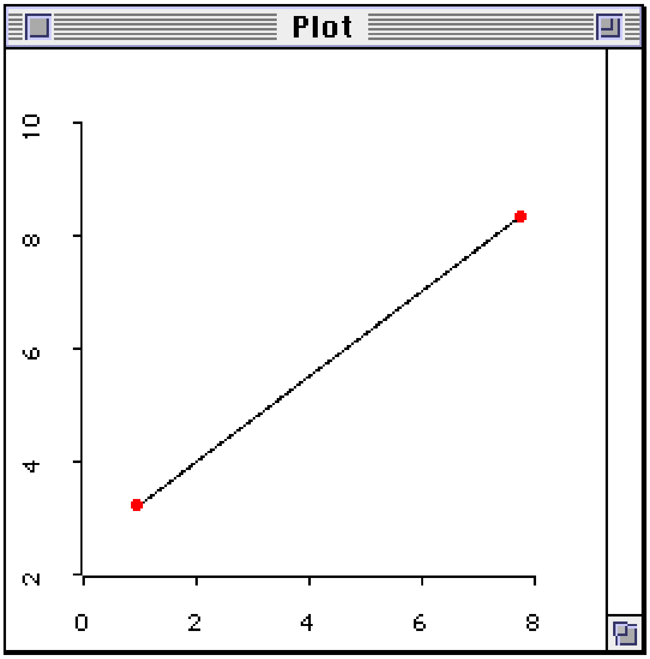
Toothaker (1986), my statistic professor at the University of Oklahoma, explain df as **the number of independent components minus the number of parameters estimated**. This approach is based upon the definition provided by Walker (1940): **the number of observation minus the number of necessary relations among these observations**.

Although Good (1973) criticized that Walker's approach is not obvious in the meaning of necessary relations, I do consider the above working definition the clearest explanation of df I ever heard. If it does not sound clear to you, I would like to use an illustration introduced by Dr. Robert Schulle, my SAS mentor at the University of Oklahoma, :

In a scatterplot when there is only one data point, you cannot do any estimation of the regression line. The line can go in any direction as shown in the following graph.



Here you have no degree of freedom (n - 1 = 0 where n = 1) for estimation. In order to plot a regession line, you must have at least two data points as indicated in the following scattergram.



In this case, you have one degree of freedom for estimation (n - 1 = 1 where n = 2). In other words, the degree of freedom tells you **the number of useful data for estimation**. However, when you have two data points only, you can always join them to be a straight regression line and get a perfect correlation (r = 1.00). Thus, the lower the degree of freedom is, the poorer the estimation is.

Mathematical Definitions

The following are indepth definitions of df:

- Good (1973) looked at degrees of freedom as the **difference of the dimensionalities of the parameter spaces**. Almost every test of a hypothesis is a test of a hypothesis H within a broader hypothesis K. Degrees of freedom, in this sense, is d(K) - d(H), where "d" stands for dimensality in parameter space.

- Galfo (1985) viewed degrees of freedom as the **representation of the quality in the given statistic which is computed using the sample X values**. Since in the computation of m, the X values can take on any of the values present in the population, the number of X values, n, selected for the given sample is the df for m. The n for the computation of m also expresses the "rung of the ladder" of quality of the m computed; i.e. if n = 1, the df, or restriction, placed on the computation is at the lowest quality level.

- Chen Xi (1994) asserted that the best way to describe the concept of the degree of freedom is in **control theory**: the degrees of freedom is **a number indicating constraints**. With the same number of the more constraints, the whole system is determined. For example, a particle moving in a three dimensional space has 9 degrees of freedom, 3 for positions, 3 for velocities, 3 for accelerations. If it is a free falling and 4 degrees of the freedom is removed, there are 2 velocities and 2 accelerations in x-y plane. There are infinite ways to add constraints, but each of the constraints will limit the moving in a certain way. The order of the state equation for a controllable and observable system is in fact the degree of the freedom.

- Cramer (1946) defined degrees of freedom as **the rank of a quadratic form**. Muirhead (1994) also adopted a **geometrical approach** to explain this concept. Degrees of freedom typically refer to chi-square distributions (and to F distributions, but they're just ratios of chi-squares). Chi-square distributed random variables are sums of squares (or quadratic forms), and can be represented as the squared lengths of vectors. The **dimension of the subspace in which the vector is free to roam is exactly the degrees of freedom**. For examples,

  - Let X_1, ... , X_n be independent N(0,1) variables, and let X be the column vector whose ith element is X_i. Then X can roam all over Euclidean n-space. Its squared length, X'X = X_1^2 + ... + X_n^2, has _a chi-square distribution with n degrees of freedom.

  - Same setup as in (1), but now let Y be the vector whose ith element is X_i-{X-bar}, where X-bar is the sample mean. Since the sum of the elements of Y must always be zero, Y cannot roam all over n-dimensional Euclidean space, but is restricted to taking values in the n-1 dimensional subspace of all n x 1 vectors whose elements sum to zero. Its squared length, Y'Y has a chi^2 distribution with n-1 degrees of freedom.

  All commonly occurring situations involving chi-sqaure distributions are similar. The most common of these are in analysis of variance (or regression) settings. F-ratios here are ratios of independent chi-square random variables, and inherit their degrees of freedom from the subspaces in which the corresponding vectors must lie.

- Rawlings (1988) associated degrees of freedom with **each sum of squares (in multiple regression) as the number of dimensions in which that vector is "free to move."** Y is free to fall anywhere in n-dimensional space and, hence, has n degrees of freedom. Y-hat, on the other hand, must fall in the X-space and , hence, has degrees of freedom equal to the dimension of the X-space -- [p', or the number of independent variables's in the model]. The residual vector e can fall anywhere in the subspace of the n-dimensional space that is orthogonal to the X-space. This subspace has dimensionality (n-p') and hence, e has (n-p') degrees of freedom.

- Selig (1994) stated that degrees of freedom are **lost for each parameter in a model that is estimated in the process of estimating another parameter**. For example, one degree of freedom is lost when we estimate the population mean using the sample mean; two degrees of freedom are lost when we estimate the standard error of estimate (in regression) using Y-hat (one degree of freedom for the Y-intercept and one degree of freedom for the slope of the regression line).

- Lambert (1994) regarded degrees of freedom as **the number of measurements exceeding the amount absolutely necessary to measure the "object" in question**. For example, to measure the diameter of a steel rod would require a minimum of one measurement. If ten measurements are taken instead, the set of ten measurements has nine degrees of freedom. In Lambert's view, once the concept is explained in this way, it is not difficult to extend it to explain applications to statistical estimators. i.e. if n measurements are made on m unknown auantities then the degrees of freedom are n-m.

Navigation

Index

Simplified Navigation

Table of Contents

Search Engine

Contact
