Degrees of Freedom

The concept of degrees of freedom is central to the principle of estimating statistics of populations from samples of them. This is a mathematical restriction that we need to put in place when we *calculate an estimate* one statistic from an *estimate* of another.

Think of data that have been drawn at random from a normal distribution. Normal distributions are characterized by only two parameters (mean and standard deviation) e.g. the standard normal distribution has a mean of 0 and standard deviation of 1. The population values of mean and standard deviation are referred to as μ and σ respectively, and the sample estimates are $\bar{x}$ and s.

In order to estimate σ, we must first have estimated µ. Thus, µ is replaced by $\bar{x}$ in the formula for σ. In other words, we work with the deviations from µ estimated by the deviations from $\bar{x}$. At this point, we need to apply the restriction that the deviations must sum to zero. Thus, degrees of freedom are *n-1* in the equation for s below:

Standard deviation in a population is:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

The estimate of population standard deviation calculated from a random sample is:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

When this principle of restriction is applied to regression and analysis of variance, the general result is that you lose one degree of freedom for each parameter estimated prior to estimating the (residual) standard deviation.

Another way of thinking about the restriction principle behind degrees of freedom is to imagine contingencies. For example, imagine you have three numbers (a, b and c) that must add up to a total of m; you are free to choose the first two numbers at random, but the third must be chosen so that it makes the total equal to m - thus your degree of freedom is two.

For example, imagine a set of three numbers, pick any number you want. For instance, it could be the set [1, 6, 5]. Calculating the mean for those numbers is easy: (1 + 6 + 5) / 3 = 4.

Now what if you imagine a set of three numbers whose mean is 3? There are lots of sets of three numbers with a mean of 3, but for any set the bottom line is this: you can freely pick the first two numbers, any number at all, but the third (last) number is out of your hands as soon as you picked the first two. Say our first two numbers are the same as in the previous set, 1 and 6, giving us a set of two freely picked numbers, and one number that we still need to choose, x: [1, 6, x]. For this set to have a mean of 3, we don’t have anything to choose about x. X has to be 2, because (1 + 6 + 2) / 3 is the only way to get to 3. So, the first two values were free for you to choose, the last value is set accordingly to get to a given mean. This set is said to have two degrees of freedom, corresponding with the number of values that you were free to choose (that is, that were allowed to vary freely).

This generalizes to a set of any given length. If I ask you to generate a set of 4, 10, or 1.000 numbers that average to 3, you can freely choose all numbers but the last one. In those sets the degrees of freedom are respectively, 3, 9, and 999. The general rule then for any set is that if *n* equals the number of values in the set, the degrees of freedom equal *n* - 1.

A data set contains a number of observations, say, *n*. They constitute *n* individual pieces of information. These pieces of information can be used either to estimate parameters or variability. In general, each item being estimated costs one degree of freedom. The remaining degrees of freedom are used to estimate variability. All we have to do is count properly.

A single sample: There are *n* observations. There's one parameter (the mean) that needs to be estimated. That leaves *n-1* degrees of freedom for estimating variability.

Two samples: There are n_1+n_2 observations. There are two means to be estimated. That leaves n_1+n_2-2 degrees of freedom for estimating variability.

Multiple regression with *p* predictors: There are *n* observations with *p+1* parameters to be estimated--one regression coeffient for each of the predictors plus the intercept. This leaves *n-p-1* degrees of freedom for error, which accounts for the error degrees of freedom in the regression.

The null hypothesis tested in the regression is that all of coefficients of the predictors are 0. The null hypothesis is that there are no coefficients to be estimated. The alternative hypothesis is that there are *p* coefficients to be estimated. Therefore, there are *p-0* or *p* degrees of freedom for testing the null hypothesis.

There is another way of viewing the Regression degrees of freedom. The null hypothesis says the expected response is the same for all values of the predictors. Therefore there is one parameter to estimate--the common response. The alternative hypothesis specifies a model with *p+1* parameters--*p* regression coefficients plus an intercept. Therefore, there are *p*--that is *p+1* (H_1) minus *1* (H_0)--regression degrees of freedom for testing the null hypothesis.

Degrees of Freedom in the t-distribution

Suppose we have a sample of *n* independent identically distributed observations drawn from a Normal population with mean μ and standard deviation σ. Let $\bar{x}$ denote the sample mean and *s*, the sample standard deviation. Then the quantity $\bar{x}$ has a *t* distribution with *n*-1 degrees of freedom. There is a different *t* distribution for each sample size. When we speak of a specific *t* distribution, we have to specify the *degrees of freedom*.

The *t- distribution* is a useful quantity when the mean and variance are unknown population parameters, in the sense that the t-value has then a probability distribution that depends on neither μ nor σ^2. The t density curves are symmetric and bell-shaped like the normal distribution and have their peak at 0. However, the spread is more than that of the standard normal distribution. This is due to the fact that the standard deviation of the population is unknown. Since s is a random quantity varying with various samples, the variability in t is more, resulting in a larger spread.

The larger the degrees of freedom, the closer the t-density is to the normal density. This reflects the fact that the standard deviation s approaches σ for large sample size n.