> *"What the hell is a degree of freedom?"*
>
> — Everyone

We keep using degrees of freedom, but its meaning can be opaque. I will try to give the intuition behind its purpose through a couple of examples.

Calculating Variance. Suppose we have $n = 4$ pieces of data, say, $\{a, b, c, d\}$. When we compute the sample mean, we must use all four pieces of data, because

$$\bar{x} = \frac{a + b + c + d}{4}.$$

Keep in mind that $\bar{x}$ is just some number.

Now consider the numerator of the sample variance,

$$(a - \bar{x})^2 + (b - \bar{x})^2 + (c - \bar{x})^2 + (d - \bar{x})^2.$$

We are still using all four pieces of data, right? *Right?*

Wrong. Because instead of using d, we could instead use

$$\bar{x} = \frac{a + b + c + d}{4} \quad \Longleftrightarrow \quad d = 4\bar{x} - a - b - c,$$

and thus could write the numerator entirely without d as

$$(a - \bar{x})^2 + (b - \bar{x})^2 + (c - \bar{x})^2 + (4\bar{x} - a - b - c - \bar{x})^2.$$

Point is, we don't need all four pieces of data to calculate the sample variance: d can be written as a dependent variable of a, b, and c, and thus we only have $3 = n - 1$ *independent* pieces of data.

Since we are therefore only really using three pieces of information, we have $3 = n - 1$ degrees of freedom, and that's why we divide the sample variance by $n - 1$ instead of n. On the other hand, when we calculated the sample mean itself, we had to use all four pieces of data, implying four degrees of freedom and thus dividing by $n = 4$.

Now let's relate this to hypothesis testing. We usually don't know the sample variance, hence we must estimate it. By the intuition above, that means we lose one degree of freedom (d becomes redundant). That is why the t-statistic has $T(n - 1)$ distribution. When we test variance itself, we have to estimate the sample variance, and hence the χ^2-statistic has $\chi^2(n - 1)$ distribution. When we calculate difference in means, we calculate sample variance for both groups, hence that t-statistic is distributed according to $T(n_1 + n_2 - 2)$.

Linear Regression. A univariate linear regression is of the form

$$y = \beta_1 + \beta_2 x + u.$$

Essentially, it is a line of best fit for the data in y and x, where u represents how far off the line is from an actual data point (because it is not a line of *perfect* fit). Since we don't know what β_1 and β_2 truly are, we have to estimate them. Our estimated regression is

$$y = b_1 + b_2 x + \widehat{u}.$$

Since b_1 and b_2 are estimates, they are subject to uncertainty, and we'd like to quantify that uncertainty. That means we'll be doing things like t-tests on them. But since we are doing estimating two parameters, b_1 and b_2, that means we'll need variances for them both. And as we know from above, a piece of information becomes redundant in the calculation of a variance. Hence, the t-statistic for, say, b_2, is

$$t = \frac{b_2 - \beta_2}{\text{se}(b_2)} \sim T(n-2).$$

Takeaway: we estimate two things, so we have $n - 2$ degrees of freedom.

If instead we have a multivariate linear regression with, say, $k = 3$ dependent variables,

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u,$$

then we'll be estimating four things: each independent variable plus the intercept. We will need variances for b_1, b_2, b_3, and b_4. Thus, we have $n - k - 1 = n - 4$ degrees of freedom. Takeaway: we're estimate four things, so we have $n - 4$ degrees of freedom.

Rule of Thumb. Every time you use a variance, you lose a degree of freedom. So if you are estimating j things, then you only have $n - j$ degrees of freedom remaining. In the context of regressions, if you have k independent variables, then you will be estimating $k+1$ variables – each regressor plus the intercept – and so you have $n-(k+1) = n-k-1$ degrees of freedom.