How to understand degrees of freedom?

From Wikipedia, there are three interpretations of the degrees of freedom of a statistic:

- In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df). In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (which, in sample variance, is one, since the sample mean is the only intermediate step).

Mathematically, degrees of freedom is the dimension of the domain of a random vector, or essentially the number of 'free' components: how many components need to be known before the vector is fully determined.

The bold words are what I don’t quite understand. If possible, some mathematical formulations will help clarify the concept.

Also do the three interpretations agree with each other?

This is a subtle question. It takes a thoughtful person not to understand those quotations! It turns out that none of them is exactly or generally correct. I haven’t the time (and there isn’t the space here) to give a full exposition, but I would like to share one approach and an insight.
Where does the concept of degrees of freedom (DF) arise? The contexts in which it’s found in elementary treatments are:

- The **Student t-test** and its variants such as the Welch or Satterthwaite solutions to the Behrens-Fisher problem (where two populations have different variances).
- The Chi-squared distribution (defined as a sum of squares of independent standard Normals), which is implicated in the **sampling distribution of the variance**.
- The **F-test** (of ratios of estimated variances).
- The **Chi-squared test**, comprising its uses in (a) testing for independence in contingency tables and (b) testing for goodness of fit of distributional estimates.

In spirit, these tests run a gamut from being exact (the Student t-test and F-test for Normal variates) to being good approximations (the Student t-test and the Welch/Satterthwaite tests for not-too-badly-skewed data) to being based on asymptotic approximations (the Chi-squared test). An interesting aspect of some of these is the appearance of non-integral "degrees of freedom" (the Welch/Satterthwaite tests and, as we will see, the Chi-squared test). This is of especial interest because it is the first hint that DF is *not* any of the things claimed of it.

**We can dispose right away of some of the claims in the question.** Because "final calculation of a statistic" is not well-defined (it apparently depends on what algorithm one uses for the calculation), it can be no more than a vague suggestion and is worth no further criticism. Similarly, neither "number of independent scores that go into the estimate" nor "the number of parameters used as intermediate steps" are well-defined.

"**Independent pieces of information that go into [an] estimate**" is difficult to deal with, because there are two different but intimately related senses of "independent" that can be relevant here. One is independence of random variables; the other is **functional independence**. As an example of the latter, suppose we collect morphometric measurements of subjects--say, for simplicity, the three dimensions $X$, $Y$, $Z$, surface areas $S = 2(XY + YZ + ZX)$, and volumes $V = XYZ$ of a set of wooden blocks. The three dimensions can be considered independent random variables, but all five variables are dependent RVs. The five are also **functionally** dependent because the **codomain (not the "domain"!)** of the vector-valued random variable $(X, Y, Z, S, V)$ traces out a three-dimensional manifold in $\mathbb{R}^5$. Thus, locally at any block $\omega$, there are two functions $f_\omega$ and $g_\omega$ for which $f_\omega(X(\psi), \ldots, V(\psi)) = 0$ and $g_\omega(X(\psi), \ldots, V(\psi)) = 0$ for blocks $\psi$ "near" $\omega$ and the derivatives of $f$ and $g$ evaluated at $\omega$ are linearly independent. However--here's the kicker--for many probability measures on the blocks, subsets of the variables such as $(X, S, V)$ are **dependent** as random variables but functionally independent.

Having been alerted by these potential ambiguities, **let's hold up the Chi-squared goodness of fit test for examination**, because (a) it’s simple, (b) it’s one of the common situations where people really do need to know about DF to get the p-value right and (c) it’s often used incorrectly. Here’s a brief synopsis of the least controversial application of this test:

- You have a collection of data values $(x_1, \ldots, x_n)$, considered an independent sample of a population.
You have estimated parameters $\theta_1, \ldots, \theta_p$ of a distribution. For example, you estimated the mean and standard deviation of a Normal distribution, expecting the population to be normally distributed.

In advance, you created a set of $k$ "bins" for the data. (It's problematic when the bins are determined by the data, even though this is often done.) Using these bins, the data are reduced to the set of counts within each bin. Anticipating what the true values of $(\theta)$ might be, you have arranged it so (hopefully) each bin will receive approximately the same count. (Equal-probability binning assures the chi-squared distribution really is a good approximation to the true distribution of the chi-squared statistic about to be described.)

You have a lot of data--enough to assure that almost all bins will have counts of 5 or greater. (This avoids small-sample problems.)

Using the parameter estimates, you can compute the expected count in each bin. The Chi-squared statistic is the sum of the ratios

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$  

This, many authorities tell us, should have (to a very close approximation) a Chi-squared distribution. But there's a whole family of such distributions, requiring a parameter $\nu$ referred to as the "degrees of freedom." The standard reasoning about how to determine $\nu$ goes like this: "I have $k$ counts. That's $k$ pieces of data. But there are (functional) relationships among them. To start with, I know in advance that the sum of the bin counts must equal $n$. That's one relationship. I estimated two (or $p$, generally) parameters from the data. That's two (or $p$) additional relationships, giving $p + 1$ total relationships. Presuming they are all (functionally) independent, that leaves only $k - p - 1$ (functionally) independent "degrees of freedom": that's the value to use for $\nu$.

The problem with this reasoning (which is exactly the sort of calculation the quotations in the question are hinting at) is that it's wrong except when some special additional conditions hold. Moreover, those conditions have nothing to do with independence (functional or statistical), with numbers of "components" of the data, with the numbers of parameters, nor with anything else referred to in the original question.

Let me show you with an example. (To make it as clear as possible, I'm using a small number of bins, but that's not essential.) Let's generate 20 iid standard Normal variates and estimate their mean and standard deviation with the usual formulas ($\text{mean} = \text{sum/count}$, etc.). To test goodness of fit, create four bins with cutpoints at the quartiles of a standard normal: -0.675, 0, +0.657, and use those counts to generate a Chi-squared statistic. Repeat as patience allows; I had time to do 10,000 repetitions.

The standard wisdom about DF says we have 4 bins and 1+2 = 3 constraints, implying the distribution of these 10,000 Chi-squared statistics should follow a Chi-squared distribution with 1 DF. Here's the histogram:
The dark blue line graphs the PDF of a $\chi^2(1)$ distribution—the one we thought would work—while the dark red line graphs that of a $\chi^2(2)$ distribution (which would be a good guess if someone were to tell you that $\nu = 1$ is incorrect). Neither fits the data.

You might expect the problem is that the data sets are small ($n=20$) or perhaps the number of bins is small. However, the problem persists even with very large datasets and larger numbers of bins: it is not merely a failure to reach an asymptotic approximation.

Things went wrong because I violated two requirements of the Chi-squared test:

1. You must use the Maximum Likelihood estimator of the parameters. (This requirement can, in practice, be slightly violated.)

2. You must base the estimate on the counts, not on the actual data! (This is crucial.)

The red histogram depicts the chi-squared statistics for 10,000 separate iterations, following these requirements. Sure enough, it visibly follows the $\chi^2(1)$ curve (with an acceptable amount of sampling error), as we had originally hoped.

The point of this comparison—which I hope you have seen coming—is that the correct DF to use for computing the p-values depends on many things other than dimensions of manifolds, counts of functional relationships, or the geometry of Normal variates. There is a subtle, delicate interaction between certain functional dependencies, as found in mathematical relationships among quantities, and distributions of the data, their statistics, and the estimators formed from them. Accordingly, it simply cannot be the case that DF is adequately explainable in terms of the geometry of multivariate normal distributions or in terms of functional independence or as counts of parameters or anything else of this nature.

We see, then, that “degrees of freedom” is merely a heuristic that suggests what the sampling
distribution of a (t, Chi-squared, or F) statistic ought to be, but it is not dispositive. Belief that it is dispositive leads to egregious errors. (For instance, the top hit on Google when searching "chi squared goodness of fit" is a Web page from an Ivy League university that gets most of this completely wrong! In particular, a simulation based on its instructions shows that the chi-squared value it recommends as having 7 DF actually has 9 DF.)

With this more nuanced understanding, it's worthwhile to re-read the Wikipedia article in question: in its details it gets things right, pointing out where the DF heuristic tends to work and where it is either an approximation or does not apply at all.

A good account of the phenomenon illustrated here (unexpectedly high DF in Chi-squared GOF tests) appears in Volume II of Kendall & Stuart, 5th edition. I am grateful for the opportunity afforded by this question to lead me back to this wonderful text, which is full of such useful analyses.

feedback

The concept is not at all difficult to make mathematical precise given a bit of general knowledge of n-dimensional Euclidean geometry, subspaces and orthogonal projections.

If P is an orthogonal projection from \( \mathbb{R}^n \) to a p-dimensional subspace \( L \) and \( x \) is an arbitrary n-vector then \( Px \) is in \( L \), \( x - Px \) and \( Px \) are orthogonal and \( x - Px \in L^\perp \) is in the orthogonal complement of \( L \). The dimension of this orthogonal complement, \( L^\perp \), is \( n - p \). If \( x \) is free to vary in an n-dimensional space then \( x - Px \) is free to vary in an \( n - p \) dimensional space. For this reason we say that \( x - Px \) has \( n - p \) degrees of freedom.

These considerations are important to statistics because if \( X \) is an n-dimensional random vector and \( L \) is a model of its mean, that is, the mean vector \( E(X) \) is in \( L \), then we call \( X - PX \) the vector of residuals, and we use the residuals to estimate the variance. The vector of residuals has \( n - p \) degrees of freedom, that is, it is constrained to a subspace of dimension \( n - p \).
If the coordinates of \( X \) are independent and normally distributed with the same variance \( \sigma^2 \) then

- The vectors \( PX \) and \( X - PX \) are independent.
- If \( E(X) \in L \) the distribution of the squared norm of the vector of residuals \( ||X - PX||^2 \) is a \( \chi^2 \)-distribution with scale parameter \( \sigma^2 \) and another parameter that happens to be the degrees of freedom \( n - p \).

The sketch of proof of these facts is given below. The two results are central for the further development of the statistical theory based on the normal distribution. Note also that this is why the \( \chi^2 \)-distribution has the parametrization it has. It is also a \( \Gamma \)-distribution with scale parameter \( 2\sigma^2 \) and shape parameter \( (n - p)/2 \), but in the context above it is natural to parametrize in terms of the degrees of freedom.

I must admit that I don't find any of the paragraphs cited from the Wikipedia article particularly enlightening, but they are not really wrong or contradictory either. They say in an imprecise, and in a general loose sense, that when we compute the estimate of the variance parameter, but do so based on residuals, we base the computation on a vector that is only free to vary in a space of dimension \( n - p \).

Beyond the theory of linear normal models the use of the concept of degrees of freedom can be confusing. It is, for instance, used in the parametrization of the \( \chi^2 \)-distribution whether or not there is a reference to anything that could have any degrees of freedom. When we consider statistical analysis of categorical data there can be some confusion about whether the "independent pieces" should be counted before or after a tabulation. Furthermore, for constraints, even for normal models, that are not subspace constraints, it is not obvious how to extend the concept of degrees of freedom. Various suggestions exist typically under the name of effective degrees of freedom.

Before any other usages and meanings of degrees of freedom is considered I will strongly recommend to become confident with it in the context of linear normal models. A reference dealing with this model class is A First Course in Linear Model Theory, and there are additional references in the preface of the book to other classical books on linear models.

**Proof of the results above:** Let \( \xi = E(X) \), note that the variance matrix is \( \sigma^2I \) and choose an orthonormal basis \( z_1, \ldots, z_p \) of \( \mathbb{L} \) and an orthonormal basis \( z_{p+1}, \ldots, z_n \) of \( \mathbb{L}^\perp \). Then \( z_1, \ldots, z_n \) is an orthonormal basis of \( \mathbb{R}^n \). Let \( \tilde{X} \) denote the \( n \)-vector of the coefficients of \( X \) in this basis, that is

\[
\tilde{X}_i = z_i^T X.
\]

This can also be written as \( \tilde{X} = Z^T X \) where \( Z \) is the orthogonal matrix with the \( z_i \)'s in the columns. Then we have to use that \( \tilde{X} \) has a normal distribution with mean \( Z^T \xi \) and, because \( Z \) is orthogonal, variance matrix \( \sigma^2I \). This follows from general linear transformation results of the normal distribution. The basis was chosen so that the coefficients of \( PX \) are \( \tilde{X}_i \) for \( i = 1, \ldots, p \), and the coefficients of \( X - PX \) are \( \tilde{X}_i \) for \( i = p + 1, \ldots, n \). Since the coefficients are uncorrelated and jointly normal, they are independent, and this implies that

\[
PX = \sum_{i=1}^p \tilde{X}_i z_i.
\]
and

\[ X - PX = \sum_{i=p+1}^{n} \tilde{X}_i z_i \]

are independent. Moreover,

\[ ||X - PX||^2 = \sum_{i=p+1}^{n} \tilde{X}_i^2. \]

If \( \xi \in L \) then \( E(\tilde{X}_i) = z_i^T \xi = 0 \) for \( i = p + 1, \ldots, n \) because then \( z_i \in L^\perp \) and hence \( z_i \perp \xi \). In this case \( ||X - PX||^2 \) is the sum of \( n - p \) independent \( N(0, \sigma^2) \)-distributed random variables, whose distribution, by definition, is a \( \chi^2 \)-distribution with scale parameter \( \sigma^2 \) and \( n - p \) degrees of freedom.
of a rectangle. Do you really know four things? No, because there are only two degrees of freedom. If you know the length and the width, you can derive the area and the perimeter. If you know the length and the area, you can derive the width and the perimeter. If you know the area and the perimeter you can derive the length and the width (up to rotation). If you have all four, you can either say that the system is consistent (all of the variables agree with each other), or inconsistent (no rectangle could actually satisfy all of the conditions). A square is a rectangle with a degree of freedom removed; if you know any side of a square or its perimeter or its area, you can derive all of the others because there's only one degree of freedom.

In statistics, things get more fuzzy, but the idea is still the same. If all of the data that you're using as the input for a function are independent variables, then you have as many degrees of freedom as you have inputs. But if they have dependence in some way, such that if you had n - k inputs you could figure out the remaining k, then you've actually only got n - k degrees of freedom. And sometimes you need to take that into account, lest you convince yourself that the data are more reliable or have more predictive power than they really do, by counting more data points than you really have independent bits of data.

Moreover, all three definitions are almost trying to give a same message.

---

Edited Oct 12 '11 at 20:52

answered Oct 12 '11 at 20:41

Biostat

632

1

10

Basically right, but I'm concerned that the middle paragraph could be read in a way that confuses correlation, independence (of random variables), and functional independence (of a manifold of parameters). The correlation-independence distinction is particularly important to maintain. -- whuber ♦ Oct 12 '11 at 20:46

@whuber: is it fine now? – Biostat Oct 12 '11 at 20:55

1 It's correct, but the way it uses terms would likely confuse some people. It still does not explicitly distinguish dependence of random variables from functional dependence. For example, the two variables in a (nondegenerate) bivariate normal distribution with nonzero correlation will be dependent (as random variables) but they still offer two degrees of freedom. -- whuber ♦ Oct 12 '11 at 21:02

biostat, thanks! I wonder if it is possible to formulate the three interpretations by Wikipedia in mathematics? That will make things clear. – Tim Oct 12 '11 at 21:22

@Tim, I believe that is done now. See my answer. – NRH Oct 12 '11 at 22:28

Feedback