$$E[Y] = E[X] + \frac{\text{Var}[X]}{E[X]}. \quad (4)$$

Remark 1: Since the variance of the exponential distribution is equal to the square of the mean, $E[Y] = 2E[X]$ when X is exponential.

Remark 2: Equation (2) yields a gamma distribution for Y when X is exponential.

Remark 3: $E[Y] \geq E[X]$, and it is only equal when X is a deterministic (and equally-spaced) interarrival time, i.e. $\text{Var}[X] = 0$.

Remark 4: Equation (4) has implications for certain survey research data that is gathered at an arbitrary point in time. When you ask a person "When did you last purchase Product A?", "When did you last attend a baseball game?", etc., you are obtaining particular outcomes of a random variable Z, where Z is the time elapsed from the last (say) purchase until the randomly selected interview time t.

In the general case (i.e., arbitrary $f(x)$) we know that $E[Z] = \frac{1}{2}E[Y]$. However, one is probably more interested in the random variable X. Obviously, one must know or hypothesize something about the functional form of the probability density for the random variable X before an estimate can be made of $E[X]$ from data on the random variable Z. For exponential interarrival times,

$$E[Z] = \tfrac{1}{2}E[Y] = E[X].$$

Since the lack of memory property applies to time in both directions, we known that Z is also distributed exponentially. $Y - Z$ will also be exponential with the same mean and will be independent of Z. Since $Y = Z + (Y - Z)$ is then the sum of two independent, identically distributed exponential random variables we further verify that Y is gamma (see Remark 2).

For deterministic (and equally-spaced) interarrival times we have

$$E[Z] = \tfrac{1}{2}E[X].$$

When only recall data are available, equation (4) may be helpful in assessing the average frequency of the performance of the recalled act.

REFERENCE

[1] Feller, William: *An Introduction to Probability Theory and its Applications*, Vol. II, Second Edition, New York: John Wiley & Sons, Inc., 1971.

What are Degrees of Freedom?

I. J. GOOD*

It might seem superfluous to be writing about the meaning of degrees of freedom half a century after Fisher's famous dispute with Karl Pearson (Fisher, 1923). But the concept is a difficult one for the student and seems to be difficult even for the expert, judging by the careless way in which it is defined in almost all books on statistics. Walker (1940) describes Tippett (1931, pp. 64–65; or see the 1952 edition, pp. 128–129) as "one of the few attempts to treat the concept of degrees of freedom in general terms, but without geometrical background." In fact Tippett's treatment is a good intuitive one for the student, but he admits himself that it is not logically complete. For a logically complete definition some reference to geometrical concepts seems essential but I do not think that even Walker's detailed geometrical discussion is entirely satisfactory. Her definition is the number of observations minus the number of necessary relations among these observations. It is not obvious what is meant by the necessary relations.

My purpose is to give a clear definition of degrees of freedom stripped of all unnecessary references, such as to normality or sums of squares.

In some contexts there is no difficulty or ambiguity. The definitions of the number of degrees of freedom in a chi-squared (tabular) distribution, and the pair of degrees of freedom in an F distribution, are clearly defined in nearly all texts. It is for *statistics*, used partly for significance testing, that obscurities arise. Most of the books resort to the device of drilling the student with techniques and rules of thumb by which he can write down the numbers of degrees of freedom.

The definition given by Cramér (1946, p. 381) is more respectable. He defines a number of degrees of freedom as the *rank* of a quadratic form. Unfortunately the rank of a quadratic form is often not obvious, and it is logically more satisfactory to frame the definition in a more general manner closely related to concepts of dimensionality.

A general definition of this type was given, for example, by Good (1967), where it was related to the Neyman-Pearson-Wilks likelihood ratio; that is, to the ratio of maximum likelihoods. It is possible, however, to formulate the definition with reference neither to likelihood ratios nor to sums of squares.

Nearly every test of a hypothesis is a test of a hypothesis H *within* a broader hypothesis K. Let the number of mathematically independent parameters in H and in K be $d(H)$ and $d(K)$ (where "d" stands for dimensionality in parameter space). *Then the number of degrees of freedom in testing* H *within* K *is* d(K) − d(H), *the difference of the dimensionalities of the parameter spaces.* This way of defining degrees of freedom does not mention the criterion by which the test is carried out. Its appropriateness is well known for the likelihood-

* Dept. of Statistics, Virginia Polytechnic Inst. & State Univ., Blacksburg, Va. 24061. This work was supported in part by H.E.W. grant No. 1 RO1 GM 18770-02.

ratio criterion in view of Wilks's result (1938) concerning the asymptotic distribution of this criterion.

It seems to me that this definition is implicit in all the best practice, and is near the surface in, for example, Lehmann (1959, Chapter 7). Unfortunately "degrees of freedom" is not listed in the index of that book as it presumably would have been had he been entirely explicit on the point.

Examples

(i) *A test for the multinomial distribution of n categories.* Here H might be a specification of the n physical probabilities, and K might leave these physical probabilities unspecified except that their total is 1. Clearly $d(H) = 0$, $d(K) = n - 1$, so d.f. $= n - 1$.

(ii) *Independence in a contingency table with r rows and s columns.* Here H asserts that the cell probability $p_{ij} = p_{i.}p_{.j} (i = 1(1)r,\ j = 1(1)s)$, where $p_{i.}$ and $p_{.j}$ denote the (unknown) marginal probabilities. The marginal probabilities are subject to two constraints, $\Sigma p_{i.} = 1$, $\Sigma p_{.j} = 1$. Hence $d(H) = r + s - 2$. Again $d(K) = rs - 1$ since $\Sigma p_{ij} = 1$, and the d.f. for testing H within K are therefore $rs - 1 - (r + s - 2) = (r - 1)(s - 1)$.

(iii) *Testing normality of N observations of a real random variable by means of an n-category (multinomial) categorization; for example, using the chi-squared test.* Here $d(H) = 2$ (because the normal distribution has two parameters) and $d(K) = n - 1$, so the number of degrees of freedom is $n - 3$.

(iv) *Additive model for an $r \times s$ two-way classification with one observation per cell.* This example is given with more elaboration by Good (1967) but we include it here (with a slight modification) for the reader's convenience. There are six natural hypotheses listed here, each being contained within the next, in parameter space. Observations x_{ij} are made corresponding to random variables $X_{ij} (i = 1(1)r,\ j = 1(1)s)$ and the hypotheses are

H_0: $X_{ij} = N(\mu + \xi_i + \eta_j,\ \sigma^2)$, a simple statistical hypothesis, that is, with all parameters specified, and with $\Sigma\xi_i = 0$, $\Sigma\eta_j = 0$. The dimensionality of the parameter space of a simple statistical hypothesis is always zero.

H_1: $X_{ij} = N(\mu + \xi_i + \eta_j,\ \sigma'^2)$, the convention being that symbols with primes denote *unspecified* parameters. Of course the standard deviation usually is unspecified so we shall not further consider hypotheses in which it is specified.

H_2: $X_{ij} = N(\mu' + \xi_i + \eta_j,\ \sigma'^2)$.

H_3: $X_{ij} = N(\xi_i' + \eta_j,\ \sigma'^2)$.

H_4: $X_{ij} = N(\xi_i' + \eta_j',\ \sigma'^2)$ ($\Sigma\xi_i' = 0$ can be forced here because any constant can be added to all the ξ_i''s if the same constant is substracted from all η_j''s.)

H_5: $X_{ij} = N(\xi_{ij}',\ \sigma'^2)$.

The dimensionalities are respectively 0, 1, 2, $r + 1$, $r + s$, and $rs + 1$ and therefore the numbers of degrees of freedom for testing H_1 within H_2, H_2 within H_3, H_3 within H_4, and H_4 within H_5 are respectively 1, $r - 1$, $s - 1$, and $(r - 1)(s - 1)$. For testing H_1 within H_5, for example, the number is rs.

The primary purpose of this note was merely to give a clear definition of degrees of freedom, not to describe its use, but perhaps a few words on this matter are in order. When the usual assumptions of normality of residuals are made, the numbers of degrees of freedom determine the appropriate chi-squared and F distributions, and the appropriate denominators of the estimates of variance components. In the more general "analysis of log-likelihood ratios" ("ANOΛ"), and in an analogous partly Bayesian analysis (Good, 1967, p. 40), the numbers of degrees of freedom are still relevant to the asymptotic forms of the corresponding distributions, as well as to their exact distributions. Closed analytic forms for these exact distributions are not known, but this lack of knowledge is not a sufficient reason for tying the meaning of degrees of freedom to normal models. The general definition of degrees of freedom is easy and natural, and encourages future analytic or numerical research into generalizations of the analysis of variance.

REFERENCES

[1] Cramér, H. (1946): *Mathematical Methods of Statistics* (Princeton Univ. Press).

[2] Fisher, R. A. (1923): "Statistical tests of agreement between observation and hypothesis", *Economica*, No. 8, pp. 1–9.

[3] Good, I. J. (1967): Contribution to the discussion of a paper by F. J. Anscombe, *Journal of the Royal Statistical Society, series B*, 29, 39–42.

[4] Lehmann, E. L. (1959): *Testing Statistical Hypotheses* (John Wiley and Sons, New York).

[5] Tippett, L. H. C. (1931): *The Methods of Statistics* (John Wiley and Sons, New York. 4th ed., 1952).

[6] Walker, Helen M. (1940): "Degrees of freedom", *Journal of Educational Psychology*, 31, 253–269. Reprinted in *Readings in Statistics for the Behavioral Scientist* (Ed. J. A. Steger; Holt, Rinehart and Winston, 1971) pp. 349–363.

[7] Wilks, S. S. (1938): "The large-sample distribution of the likelihood-ratio for testing composite hypotheses", *Annals of Mathematical Statistics*, 9, 60–62.