Table 8.4.2.1. *Values of the F ratio for which the c.d.f. $\Psi(F,\nu_1,\nu_2)$ has the value 0.95, for various choices of ν_1 and ν_2*

ν_2 \ ν_1	1	2	4	8	15
10	4.9646	4.1028	3.4781	3.0717	2.8450
20	4.3512	3.4928	2.8661	2.4471	2.2033
30	4.1709	3.3158	2.6896	2.2662	2.0148
40	4.0847	3.2317	2.6060	2.1802	1.9245
50	4.0343	3.1826	2.5572	2.1299	1.8714
60	4.0012	3.1504	2.5252	2.0970	1.8364
80	3.9604	3.1108	2.4859	2.0564	1.7932
100	3.9361	3.0873	2.4626	2.0323	1.7675
120	3.9201	3.0718	2.4472	2.0164	1.7505
150	3.9042	3.0564	2.4320	2.0006	1.7335
200	3.8884	3.0411	2.4168	1.9849	1.7167
300	3.8726	3.0259	2.4017	1.9693	1.6998
400	3.8648	3.0183	2.3943	1.9616	1.6914
600	3.8570	3.0107	2.3868	1.9538	1.6831
1000	3.8508	3.0047	2.3808	1.9477	1.6764

The marginal p.d.f. for F is obtained by integration of the joint p.d.f.,

$$\Phi(F) = \int_0^\infty \Phi_C\left(F|\chi_2^2\right)\Phi_M\left(\chi_2^2\right)\,\mathrm{d}\chi_2^2, \tag{8.4.2.3}$$

yielding the result

$$\Phi(F,\nu_1,\nu_2) = \frac{(\nu_1/\nu_2)F^{\nu_1/2-1}}{B(\nu_1/2,\nu_2/2)\left[1+(\nu_1/\nu_2)F\right]^{(\nu_1+\nu_2)/2}}. \tag{8.4.2.4}$$

This p.d.f. is known as the *F distribution with ν_1 and ν_2 degrees of freedom*. Table 8.4.2.1 gives the values of F for which the c.d.f. $\Psi(F,\nu_1,\nu_2)$ is equal to 0.95 for various choices of ν_1 and ν_2. Fortran code for the program from which the table was generated appears in Prince (1994).

The cumulative distribution function $\Psi(F,\nu_1,\nu_2)$ gives the probability that the F ratio will be less than some value by chance if the models are equally consistent with the data. It is therefore a necessary, but not sufficient, condition for concluding that the unconstrained model gives a significantly better fit to the data that $\Psi(F,\nu_1,\nu_2)$ be greater than $1-\alpha$, where α is the desired level of significance. For example, if $\Psi(F,\nu_1,\nu_2)=0.95$, the probability is only 0.05 that a value of F this large or greater would have been observed if the two models were equally good representations of the data.

Hamilton (1964) observed that the F ratio could be expressed in terms of the crystallographic weighted R index, which is defined, for refinement on |F| (and similarly for refinement on $|\mathrm{F}|^2$), by

$$R_w = \left[\sum w_i(|\mathrm{F}_o|_i - |\mathrm{F}_c|_i)^2 \Big/ \sum w_i|\mathrm{F}_o|_i^2\right]^{1/2}. \tag{8.4.2.5}$$

Denoting by R_c and R_u the weighted R indices for the constrained and unconstrained models, respectively,

$$F = (\nu_2/\nu_1)[(R_c/R_u)^2 - 1], \tag{8.4.2.6}$$

and a c.d.f. for R_c/R_u can be readily derived from this relation. A significance test based on R_c/R_u is known as *Hamilton's R-ratio test*; it is entirely equivalent to a test on the F ratio.

8.4.3. Comparison of different models

Tests based on F or the R ratio have several limitations. One important one is that they are applicable only when the parameters of one model form a subset of the parameters of the other. Also, the F test makes no distinction between improvement in fit as a result of small improvements throughout the entire data set and a large improvement in a small number of critically sensitive data points. A test that can be used for comparing arbitrary pairs of models, and that focuses attention on those data points that are most sensitive to differences in the models, was introduced by Williams & Kloot (1953; also Himmelblau, 1970; Prince, 1982).

Table 8.4.3.1. *Values of t for which the c.d.f. $\Psi(t,\nu)$ has the values given in the column headings, for various values of ν*

ν	0.75	0.90	0.95	0.99	0.995
1	1.0000	3.0777	6.3138	31.8206	63.6570
2	0.8165	1.8856	2.9200	6.9646	9.9249
3	0.7649	1.6377	2.3534	4.5407	5.8409
4	0.7407	1.5332	2.1319	3.7469	4.6041
6	0.7176	1.4398	1.9432	3.1427	3.7074
8	0.7064	1.3968	1.8596	2.8965	3.3554
10	0.6998	1.3722	1.8125	2.7638	3.1693
12	0.6955	1.3562	1.7823	2.6810	3.0546
14	0.6924	1.3450	1.7613	2.6245	2.9769
16	0.6901	1.3368	1.7459	2.5835	2.9208
20	0.6870	1.3253	1.7247	2.5280	2.8453
25	0.6844	1.3164	1.7081	2.4851	2.7874
30	0.6828	1.3104	1.6973	2.4573	2.7500
35	0.6816	1.3062	1.6896	2.4377	2.7238
40	0.6807	1.3031	1.6839	2.4233	2.7045
50	0.6794	1.2987	1.6759	2.4033	2.6778
60	0.6786	1.2958	1.6707	2.3901	2.6603
80	0.6776	1.2922	1.6641	2.3739	2.6387
100	0.6770	1.2901	1.6602	2.3642	2.6259
120	0.6765	1.2886	1.6577	2.3578	2.6174

Consider a set of observations, y_{0i}, and two models that predict values for these observations, y_{1i} and y_{2i}, respectively. We determine the slope of the regression line $z=\lambda x$, where $z_i = \left[y_{0i} - (1/2)(y_{1i}+y_{2i})\right]/\sigma_i$, and $x_i = (y_{1i}-y_{2i})/\sigma_i$. Suppose model 1 is a perfect fit to the data, which have been measured with great precision, so that $y_{0i}=y_{1i}$ for all i. Under these conditions, $\lambda=+1/2$. Similarly, if model 2 is a perfect fit, $\lambda=-1/2$. Real experimental data, of course, are subject to random error, and $|\lambda|$ in general would be expected to be less than 1/2. A least-squares estimate of λ is

$$\widehat{\lambda} = \frac{\sum_{i=1}^n z_i x_i}{\sum_{i=1}^n x_i^2}, \tag{8.4.3.1}$$

and it has an estimated variance

$$\widehat{\sigma}_\lambda^2 = \frac{\sum_{i=1}^n z_i^2 - \widehat{\lambda}^2\sum_{i=1}^n x_i^2}{(n-1)\sum_{i=1}^n x_i^2}. \tag{8.4.3.2}$$

The hypothesis that the two models give equally good fits to the data can be tested by considering $\widehat{\lambda}$ to be an unconstrained, one-parameter fit that is to be compared with a constrained, zero-parameter fit for which $\lambda=0$. A p.d.f. for making this comparison can be derived from an F distribution with $\nu_1=1$ and $\nu_2=\nu=(n-1)$.

$$\Phi(F,1,\nu) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu F}\,\Gamma(\nu/2)(1+F/\nu)^{(\nu+1)/2}}. \tag{8.4.3.3}$$

If we let $|t| = \sqrt{F}$, and use

$$\int_0^{F_0} \Phi(F, 1, \nu)\, \mathrm{d}F = \int_{-t_0}^{+t_0} \Phi(t, \nu)\, \mathrm{d}t, \tag{8.4.3.4}$$

we can derive a p.d.f. for t, which is

$$\Phi(t, \nu) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)[1 + t^2/\nu]^{(\nu+1)/2}}. \tag{8.4.3.5}$$

This p.d.f. is known as *Student's t distribution with ν degrees of freedom*. Setting $t = \widehat{\lambda}/\widehat{\sigma}_\lambda$, the c.d.f. $\Psi(t, \nu)$ can be used to test the alternative hypotheses $\lambda = 0$ and $\lambda = \pm 1/2$. Table 8.4.3.1 gives the values of t for which the c.d.f. $\Psi(t, \nu)$ has various values for various values of ν. Fortran code for the program from which this table was generated appears in Prince (1994).

Again, it must be understood that the results of these statistical comparisons do not imply that either model is a correct one. A statistical indication of a good fit says only that, given the model, the experimenter should not be surprised at having observed the data values that were observed. It says nothing about whether the model is plausible in terms of compatibility with the laws of physics and chemistry. Nor does it rule out the existence of other models that describe the data as well as or better than any of the models tested.

8.4.4. Influence of individual data points

When the method of least squares, or any variant of it, is used to refine a crystal structure, it is implicitly assumed that a model with adjustable parameters makes an unbiased prediction of the experimental observations for some (*a priori* unknown) set of values of those parameters. The existence of any reflection whose observed intensity is inconsistent with this assumption, that is that it differs from the predicted value by an amount that cannot be reconciled with the precision of the measurement, must cause the model to be rejected, or at least modified. In making precise estimates of the values of the unknown parameters, however, different reflections do not all carry the same amount of information (Shoemaker, 1968; Prince & Nicholson, 1985). For an obvious example, consider a space-group systematic absence. Except for possible effects of multiple diffraction or twinning, any observed intensity at a position corresponding to a systematic absence is proof that the screw axis or glide plane is not present. If no intensity is observed for any such reflection, however, any parameter values that conform to the space group are equally acceptable. It is to be expected, on the other hand, that some intensities will be extremely sensitive to small changes in some parameter, and that careful measurement of those intensities will lead to correspondingly precise estimates of the parameter values. For the purpose of precise structure refinement, it is useful to be able to identify the influential reflections.

Consider a vector of observations, $\mathbf{y}$, and a model $\boldsymbol{M}(\mathbf{x})$. The elements of $\mathbf{y}$ define an n-dimension space, and the model values, $M_i(\mathbf{x})$, define a p-dimensional subspace within it. The least-squares solution [equation (8.1.2.7)],

$$\widehat{\mathbf{x}} = (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}(\mathbf{y} - \mathbf{y}_0), \tag{8.4.4.1}$$

is such that $\widehat{\mathbf{y}} = \boldsymbol{M}(\widehat{\mathbf{x}})$ is the closest point to $\mathbf{y}$ that corresponds to some possible value of $\mathbf{x}$. In (8.4.4.1), $\boldsymbol{W} = \boldsymbol{V}^{-1}$ is the inverse of the variance–covariance matrix for the joint p.d.f. of the elements of $\mathbf{y}$, and $\mathbf{y}_0 = \boldsymbol{M}(\mathbf{x}_0)$ is a point in the p-dimensional subspace close enough to $\boldsymbol{M}(\widehat{\mathbf{x}})$ so that the linear approximation

$$\boldsymbol{M}(\mathbf{x}) = \mathbf{y}_0 + \boldsymbol{A}(\mathbf{x} - \mathbf{x}_0) \tag{8.4.4.2}$$

[where $A_{ij} = \partial M_i(\mathbf{x})/\partial x_j$] is a good one. Let $\boldsymbol{R}$ be the Cholesky factor of $\boldsymbol{W}$, so that $\boldsymbol{W} = \boldsymbol{R}^T\boldsymbol{R}$, and let $\boldsymbol{Z} = \boldsymbol{R}\boldsymbol{A}$, $\mathbf{y}' = \mathbf{y} - \mathbf{y}_0$, and $\widehat{\mathbf{y}}' = \widehat{\mathbf{y}} - \mathbf{y}_0$. The least-squares estimate may then be written

$$\widehat{\mathbf{x}} = \mathbf{x}_0 + (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{y}' \tag{8.4.4.3}$$

and

$$\widehat{\mathbf{y}}' = \boldsymbol{Z}(\widehat{\mathbf{x}} - \mathbf{x}_0) = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{y}'. \tag{8.4.4.4}$$

Thus, the matrix $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})\boldsymbol{Z}^T$, the *projection matrix*, is a linear relation between the observed data values and the corresponding calculated values. (Because $\widehat{\mathbf{y}}' = \boldsymbol{P}\mathbf{y}'$, the matrix $\boldsymbol{P}$ is frequently referred to in the statistical literature as the *hat matrix*.) $\boldsymbol{P}^2 = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T = \boldsymbol{P}$, so that $\boldsymbol{P}$ is *idempotent*. $\boldsymbol{P}$ is an $n \times n$ positive semidefinite matrix with rank p, and its eigenvalues are either 1 (p times) or 0 ($n - p$ times). Its diagonal elements lie in the range $0 \leq P_{ii} \leq 1$, and the trace of $\boldsymbol{P}$ is p, so that the average value of P_{ii} is p/n. Furthermore,

$$P_{ii} = \sum_{j=1}^{n} P_{ij}^2. \tag{8.4.4.5}$$

A diagonal element of $\boldsymbol{P}$ is a measure of the influence that an observation has on its own calculated value. If P_{ii} is close to one, the model is forced to fit the ith data point, which puts a constraint on the value of the corresponding function of the parameters. A very small value of P_{ii}, because of (8.4.4.5), implies that all elements of the row must be small, and that observation has little influence on its own or any other calculated value. Because it is a measure of influence on the fit, P_{ii} is sometimes referred to as the *leverage* of the ith observation. Note that, because $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \boldsymbol{V}_{\mathbf{x}}$, the variance–covariance matrix for the elements of $\widehat{\mathbf{x}}$, $\boldsymbol{P}$ is the variance–covariance matrix for $\widehat{\mathbf{y}}$, whose elements are functions of the elements of $\widehat{\mathbf{x}}$. A large value of P_{ii} means that y_i is poorly defined by the elements of $\widehat{\mathbf{x}}$, which implies in turn that some elements of $\widehat{\mathbf{x}}$ must be precisely defined by a precise measurement of y_i'.

It is apparent that, in a real experiment, there will be appreciable variation among observations in their leverage. It can be shown (Fedorov, 1972; Prince & Nicholson, 1985) that the observations with the greatest leverage also have the largest effect on the volume of the p-dimensional confidence region for the parameter estimates. Because this volume is a rather gross measure, however, it is useful to have a measure of the influence of individual observations on individual parameters. Let $\boldsymbol{V}_n$ be the variance–covariance matrix for a refinement including n observations, and let $\mathbf{z}$ be a row vector whose elements are $z_j = [\partial M(\mathbf{x})/\partial x_j]/\sigma$ for an additional observation. $\boldsymbol{V}_{n+1}$, the variance–covariance matrix with the additional observation included, is, by definition,

$$\boldsymbol{V}_{n+1} = (\boldsymbol{Z}^T\boldsymbol{Z} + \mathbf{z}^T\mathbf{z})^{-1}, \tag{8.4.4.6}$$

which, in the linear approximation, can be shown to be

$$\boldsymbol{V}_{n+1} = \boldsymbol{V}_n - \boldsymbol{V}_n\mathbf{z}^T\mathbf{z}\boldsymbol{V}_n/(1 + \mathbf{z}\boldsymbol{V}_n\mathbf{z}^T). \tag{8.4.4.7}$$

The diagonal elements of the rank one matrix $\boldsymbol{D} = \boldsymbol{V}_n\mathbf{z}^T\mathbf{z}\boldsymbol{V}_n/(1 + \mathbf{z}\boldsymbol{V}_n\mathbf{z}^T)$ are therefore the amounts that the variances of the estimates of individual parameters will be reduced by inclusion of the additional observation.

This result depends on the elements of $\boldsymbol{Z}$ and $\mathbf{z}$ not changing significantly in the (presumably small) shift from $\widehat{\mathbf{x}}_n$ to $\widehat{\mathbf{x}}_{n+1}$. That this condition is satisfied may be verified by the following procedure. Find an approximation to $\widehat{\mathbf{x}}_{n+1}$ by a line search

references