

7. Hypothesis Testing

- *the Chisquare distribution*
- *the F test*
- *multiple-regression analysis*

7. Hypothesis testing

a. Purpose

Chapters 4 and 5 discussed methods for determining the best-fit values of parameters, but did not consider if those values provided a good fit to the data. However, it is equally necessary to consider if the selected representation of the data is adequate or consistent, and if the number of parameters is sufficient to represent the observations. The statistical basis for addressing these questions will only be treated in cursory fashion here. This is a branch of mathematics that has received extensive application in meteorology, especially in connection with weather modification experiments; see, for example, Dennis (19XX).

Here the discussion is limited to the chisquare test, the F test, and a short warning concerning likelihood ratios. The chisquare test and the Student t test (discussed in Chapter 2) are the tests most often needed in analyses of experimental data. The F test is included here because, among other applications, it is useful when considering how many parameters must be included in multiple-parameter fits. However, it is expected that these sections are only short reminders of familiar material. If not, the reader will benefit from study of a more complete text, such as one of those included in the bibliography.

b. *The Chisquare distribution*

Section 4d developed the connection between the method of least squares and the maximum-likelihood method, and showed that the one-standard-deviation limits for the fit correspond to changes in fitted parameters that increase the chisquare by one. Consider, for example, the chisquare function for the mean u of a set of measurements:

$$\chi^2(u) = \sum_i \frac{(y_i - u)^2}{\sigma_i^2} \quad (7.1)$$

$$\begin{aligned} \chi^2(u \pm \sigma_u) &= \sum_i \frac{(y_i - (u \pm \sigma_u))^2}{\sigma_i^2} \\ &= \sum_i \frac{(y_i - u)^2 \pm 2\sigma_u(y_i - u) + \sigma_u^2}{\sigma_i^2} \\ &= \chi^2(u) + \sigma_u^2 \sum_i \frac{1}{\sigma_i^2}. \end{aligned} \quad (7.2)$$

Because

$$\sigma_u^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}, \quad (7.3)$$

$$\chi^2(u \pm \sigma_u) = \chi^2(u) + 1, \quad (7.4)$$

the same result obtained more generally in section 5c.

It is worth re-emphasis that this result depends on the validity of the expected Gaussian distribution of errors, and has the same connection to this assumption as does the least-squares method of fitting to data. The basic equation used here for the chisquare function, (7.1), depends on this assumption, and (7.3) is only valid if the errors from individual measurements entering the mean are uncorrelated. Inference based on the chisquare distribution will not be valid if these conditions are not satisfied.

It is useful to consider the distribution in values expected for the chisquare function when these conditions are satisfied. Consider the case where the correct functional form $f(x)$ is used in a fit to measurements $\{y_i\}$ obtained at values $\{x_i\}$ of the independent variable x . If there are N measurements in the fit, each characterized by the same measurement uncertainty σ , the variance of the measurements about the best-fit relationship is

$$V = \frac{1}{N} \sum_i (y_i - f(x_i))^2 = \frac{1}{N} \sigma^2 \chi^2. \quad (7.5)$$

If there are n parameters in the fit,

$$\langle s^2 \rangle = \sigma^2 \frac{N}{N - n} \quad (7.6)$$

when $f(x)$ is the correct functional relationship, so

$$\langle \chi^2 \rangle = N - n = \text{degrees of freedom.} \quad (7.7)$$

For example, for 25 measurements and a fit with three parameters, $\chi^2 \approx 22$ is expected if the functional relationship is correct.

For the chisquare distribution function, the expected value and the distribution about that expected value both depend on the number of degrees of freedom. The distribution is that expected for the sum of the squares of $\nu = N - n$ independent unit-normal variables. The functional form of the chisquare distribution is¹

$$P(z, \nu) = \frac{1}{2\Gamma(\nu/2)} \left(\frac{z}{2}\right)^{\frac{\nu}{2}-1} e^{-z/2} \quad (7.8)$$

where ν is the number of degrees of freedom and z is the value of the chisquare function. Γ is the generalized factorial function, defined so that $\Gamma(1) = 1$ and $\Gamma(n+1) = n\Gamma(n)$. (For integer n , $\Gamma(n+1) = n!$, and $\Gamma(1/2) = \sqrt{\pi}$.)

Figure 7.1 shows the probability that χ^2 will exceed various limits, as a function of the number of degrees of freedom. These curves make it possible to judge how consistent a fit is with the data. If a chisquare is obtained that corresponds to a very unlikely value (e.g., if only about 5% of all observations are expected to have this large a chisquare), then the fit is not very good and the functional dependence is probably not correct.

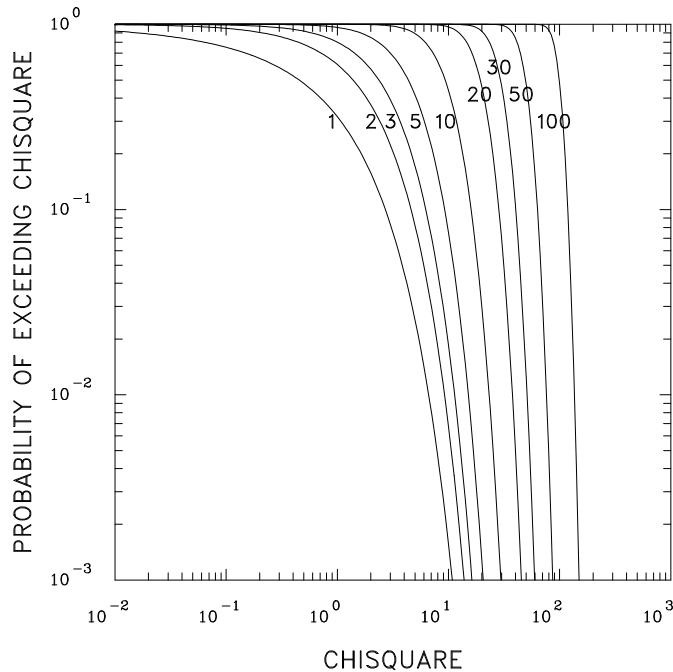


Fig. 7.1: Probability of exceeding various values of the chisquare, for the degrees of freedom labeled on the curves.

The following properties of the chisquare distribution function are often useful:

1. $\langle \chi^2 \rangle = \nu$, as shown before, where ν is the number of degrees of freedom in the fit.
2. The variance in the value of χ^2 is $V(\chi^2) = 2\nu$.
3. The chisquare function approaches a Gaussian distribution with the above mean and variance, for large ν .
4. The most probable value in the distribution occurs for $\chi^2 = \nu - 2$.

¹ Cf., e.g., Brownlee 1965 for a derivation

One valuable use of the chisquare statistic is in considerations of how many parameters are needed to account for the observed variability. If the chisquare test indicates an unsatisfactory fit, it may be necessary to add additional parameters to the fit.

c. The F-test

If $f(x)$ is used to approximate measurements $\{y_i\}$, and if the number of degrees of freedom is ν , the sample estimate of the standard deviation is related to the chisquare by

$$s^2 = \frac{1}{\nu} \sum_i (y_i - f(x_i))^2 = \frac{\sigma^2}{\nu} \chi^2. \quad (7.9)$$

Consider two samples taken from the same population, both characterized by the same standard deviation σ . Define

$$F = \frac{s_1^2}{s_2^2} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}. \quad (7.10)$$

The distribution function for F can be derived as the ratio of two chisquare distribution functions. It is:

$$P(F, \nu_1, \nu_2) = \frac{\Gamma[(\nu_1 + \nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{F^{\frac{1}{2}(\nu_1-1)}}{(1 + F \frac{\nu_1}{\nu_2})^{\frac{\nu_1+\nu_2}{2}}} \quad (7.11)$$

An approximation that is usually adequate is to use the following variable as a normal deviate:

$$z = \frac{F^{\frac{1}{3}}(1 - \frac{2}{9\nu_2}) - (1 - \frac{2}{9\nu_1})}{(\frac{2}{9\nu_1} + F^{2/3} \frac{2}{9\nu_2})^{1/2}}. \quad (7.12)$$

The F-test can be used to determine if two samples are consistent with a common origin. It is used to compare the sample variances, as follows. Consider an example where there are two sets of measurements to be tested for consistency, one with 6 degrees of freedom and a sample estimate of variance of $s_1^2=75$ and a second with 10 degrees of freedom and a sample estimate of variance of $s_2^2=25$. To determine if the two samples are different at the 90% confidence level:

1. $F = (s_1^2/s_2^2) = 3$ with $f_1=6$ and $f_2=10$.
2. For a 90% confidence test, use a 5% test for both the upper and lower tails of the distribution.
3. Reference tables² show 3.22 to be the critical value of F for a 5% confidence interval. $F=3.00$ is thus less than this critical value, so the difference is not significant at the 5% level.
4. It is also necessary to test if the ratio is too small. The 95% limit for the same ratio $F(s_1^2/s_2^2)$ can be found by using the symmetry in the tables because the 95% limit for $f_1=6$ and $f_2=10$ is the inverse of the 5% limit for $f_1=10$ and $f_2=6$, so the lower limit is $1/4.08=0.245$. The value 3.0 is well above this lower limit.

Thus the samples, while apparently quite different, do not fail a 90% confidence test that they are the same. It would be a serious misinterpretation of this test to conclude from

² e.g., Abramowitz, M. and I. A. Stegun, *Handbook of Mathematical Functions*, 1970, Dover Publications, New York, p. 987

these results that they *are* the same; the correct conclusion is that the hypothesis that they are the same cannot be rejected with 90% confidence. Indeed, the test will fail at about the 87% level, or alternately a one-sided test (applicable if the direction of the difference between the samples is prescribed in advance) will fail at about the 94% level, so there is a strong indication that the two samples are different even though the posed hypothesis cannot be rejected at the 90% confidence level.

When the Gaussian approximation is used, the two test values for z are $z=1.550$ and $z'=-1.550$. These values correspond to the 93.9% and 6.1% cumulative points in the Gaussian distribution, so the test would fail a test with about an 88% confidence limit although it passes the 90% test. The accuracy of this approximation is demonstrated by evaluating z for $F=3.22$, $f_1=6$ and $f_2=10$, which gives $z=1.645$, a value corresponding to the 0.9500 point in the cumulative Gaussian distribution function. The Gaussian approximation is thus very accurate in this case, and is almost always acceptable.

d. Use of the F test in multiple-regression analysis

Consider the case of multiple linear regression, where the function $f(x)$ with linear parameters $\{b_k\}$ is used to fit to a set of measurements $\{y_i\}$:

$$f(x_i) = \bar{y} + \sum_k b_k (X_{ki} - \bar{X}_k) \quad (7.13)$$

where $X_{ki}=g_k(x_i)$ is the value of the k -th function of x evaluated at x_i . The chisquare for a fit to this function is then

$$\chi^2 = \sum_i \frac{(y_i - \bar{y} - \sum_k b_k (X_{ki} - \bar{X}_k))^2}{\sigma_i^2}. \quad (7.14)$$

The least-squares solution is

$$\frac{\partial \chi^2}{\partial b_\ell} = -2 \sum_i \frac{y_i - \bar{y} - \sum_k b_k (X_{ki} - \bar{X}_k) (X_{\ell i} - \bar{X}_\ell)}{\sigma_i^2} = 0 \quad (7.15)$$

or

$$\sum_i \frac{(y_i - \bar{y})(X_{\ell i} - \bar{X}_\ell)}{\sigma_i^2} = \sum_i \sum_k \frac{b_k (X_{ki} - \bar{X}_k) (X_{\ell i} - \bar{X}_\ell)}{\sigma_i^2}. \quad (7.16)$$

Then, at this best-fit solution,

$$\chi^2 = \sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2} - \sum_i \frac{(y_i - \bar{y}) \sum_k b_k (X_{ki} - \bar{X}_k)}{\sigma_i^2}. \quad (7.17)$$

In the case of simple linear regression with one parameter,

$$\chi^2 = \sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2} - \sum_i \frac{(y_i - \bar{y}) b (x_i - \bar{x})}{\sigma_i^2} \quad (7.18)$$

or, using the earlier results (6.3) and (6.9) for the regression slope b and correlation coefficient r ,

$$\chi^2 = \sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2} (1 - r^2). \quad (7.19)$$

This result provides a basis for extending the correlation coefficient to the case of multiple regression. Define the multiple regression coefficient R so that it will have an analogous role to r in (7.19):

$$\begin{aligned} R^2 &= \frac{\sum_k b_k \sum_i \frac{(y_i - \bar{y})(X_{ki} - \bar{X}_k)}{\sigma_i^2}}{\sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2}} \\ &= \sum_k \frac{b_k V_{y, X_k}}{V_{yy}} = \sum_k b_k \left(\frac{V_{X_k X_k}}{V_{yy}} \right)^{1/2} r_{X_k y} = \sum_k b_k \frac{\sigma_{X_k}}{\sigma_y} r_{X_k y}. \end{aligned} \quad (7.20)$$

Then, as desired,

$$\chi^2 = \sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2} (1 - R^2). \quad (7.21)$$

If the standard deviations σ_i are known, a chisquare test can determine if the multiple-regression fit is an adequate representation of the data. An alternate test that does not require knowledge of the true standard deviations is to form the ratio

$$F = \frac{\chi^2(n-1) - \chi^2(n)}{\chi^2(n)/(N-n)} \quad (7.22)$$

where N is the number of data values used in the fit and n the number of parameters. The numerator in this equation is the difference between the chisquare calculated for n parameters and that for $n-1$ parameters; it is therefore distributed as a chisquare function with 1 degree of freedom and has expected value 1 if the fit does not improve significantly on addition of the new parameter.

If the fits with $(n-1)$ and with n parameters are both as good as expected for correct functional relationships, we expect $\chi^2(n)$ to be $(N-n-1)$ and F to be 1. Small values of F (near 1) then indicate that the fit has not improved significantly; the improvement in the fit by addition of a parameter is consistent with the improvement expected because the degrees of freedom are reduced. If, on the other hand, F is much larger than 1, the fit has improved more than would be expected from addition of another non-physical parameter, so the added parameter provides a useful improvement in the fit by representing a true variation in the data.

If it is planned to test 20 parameters for inclusion in a multiple-regression analysis, a typical choice might be to require that the F-test for addition of each accepted parameter have a significance level less than 0.05, so that an average of only one insignificant parameter will be included in the fit. Another approach is to add the most significant parameter from the set at each step, and continue until an acceptable representation of the data or acceptable multiple-correlation coefficient is obtained.

It is also possible to use the multiple correlation coefficient itself as a test. If there are n regression variables (giving $n+1$ parameters, including a constant), there are $N-n-1$ degrees of freedom, and

$$\sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2}$$

will be characterized by $(N - 1)$ degrees of freedom. Then the ratio

$$F = \frac{\sum \frac{(y_i - \bar{y})^2}{\sigma_i^2} / ((N - 1))}{(1 - R^2) \sum_i \frac{(y_i - \bar{y})^2}{\sigma_i^2} / (N - n - 1)} = \frac{N - n - 1}{(N - 1)(1 - R^2)} \quad (7.23)$$

will be distributed according to an F distribution with $(N - 1, N - n - 1)$ degrees of freedom. A value of F near 1 thus indicates that the composite regression does not determine a significant functional relationship; even with no true correlation, it is expected that R^2 will differ from 0 as n increases. For example, with 100 points and 50 parameters, R^2 will be about 0.5 even if there are no true correlations in the data. If those 50 parameters were considered individually, on the average one of them would show a significant correlation with the measurements at a confidence limit of 0.02 (for $F \approx 1.5$). Because of these results, searches through large numbers of parameters for significant relationships can give misleading results.

A convenient way to test the significance of added terms in a regression analysis is as follows. Construct an F -test statistic to test the hypothesis that the added term is *not* significant. If the term is not significant, the sample estimate of the variance about the best-fit relationship will not reduce when the new parameter is added to the fit. A useful test statistic is then

$$F = \frac{\chi_{n-1}^2 / (N - n - 2)}{\chi_n^2 / (N - n - 1)} = \frac{(1 - R_{(n-1)}^2)(N - n - 1)}{(1 - R_n^2)(N - n - 2)}. \quad (7.24)$$

As an example, suppose that $N=100$, $n=5$, and $R=0.9$. When a sixth parameter is added, R becomes 0.93. Is the addition significant?

For this case, $F=1.421$, $z^*=1.69$, and $P(z \geq z^*) \approx 0.05$, so the added parameter provides an improvement to the fit that is significant at the 5% level.

e. Hypothesis testing by likelihood ratios

In cases where statistical inference is difficult or unwieldy by other means, likelihood ratios often prove useful because of their general applicability. A likelihood ratio can be defined as

$$\lambda = \frac{L(A)}{L(B)} \quad (7.25)$$

where A represents the multidimensional volume corresponding to a particular hypothesis and B represents the union of volumes corresponding to that hypothesis and its alternative. In the case where there are n parameters, the hypothesis might be that $x_1 = c$ with no restriction on other parameters. To test the hypothesis:

1. Calculate $L(A)$, the maximum likelihood with $x_1 = c$. That is, find the set of parameters $\{x_i\}$, $i \neq 1$, that give the maximum likelihood while x_1 is constrained to be c .
2. Calculate $L(B)$, the usual maximum likelihood without any restriction on x_1 .
3. Calculate λ from (7.25).
4. A value of λ near unity indicates that the hypothesis should be accepted, while a small value of λ suggests rejection.

- Specifically, for large sample sizes, the parameter $v = -2\ln(\lambda)$ will be distributed as χ^2 for $f_B - f_A$ degrees of freedom, where $f_B(f_A)$ is the number of degrees of freedom for hypothesis B (A).

Tests based on likelihood ratios are notoriously difficult to interpret, and are often misleading. The reason is that, if the functional form is wrong or the fit is poor, large likelihood ratios will result, and the resulting large changes in values of the likelihood when parameters change by small amounts can lead an analyst to think that the best-fit values are determined with great accuracy when instead the correct interpretation is that the fit is not adequate to represent the data. Confidence limits should be determined from likelihood ratios only in those cases where there is evidence that the fit is adequate.

SOURCES AND FURTHER READING

- Abramowitz, M. and I. A. Stegun, 1972: Handbook of Mathematical Functions. Dover Publications, New York, 1046 pp.
- Anderson, V. L., and R. A. McLean, 1974: Design of Experiments. Marcel Dekker, Inc., New York, 418 pp.
- Box, G. E. P., W. G. Hunter, and J. S. Hunter, 1978: Statistics for Experimenters. John Wiley and Sons, New York, 653 pp.
- Brownlee, K. A., 1965: Statistical Theory and Methodology in Science and Engineering. John Wiley and Sons, New York, 590 pp.
- Murphy, A. H., and R. W. Katz, 1985: Probability, Statistics, and Decision Making in the Atmospheric Sciences. Westview Press, Boulder, Colorado, 545 pp.
- Panofsky, H. A., and G. W. Brier, 1968: Some applications of Statistics to Meteorology. Pennsylvania State University, 224 pp.