## Development of the General Form of the Test Statistic

In the following paragraphs, we develop this idea of proportional increase in error into a test statistic. Our development does not proceed in the way the test statistic would be introduced in a mathematical statistics text. However, our goal is like the mathematician's in that we strive for generality, not just the solution to a single problem. We develop the test statistic rationally, not mathematically, as a reasonable index of the relative adequacy yet simplicity of two competing models. However, instead of developing things in a way that would work only in a one-sample situation, we introduce a method that works in essentially all cases we consider in this book. Doing so takes a few more lines than developing a test for only one sample. However, in so doing, we are providing a perspective and a general procedure that together serve as a unifying theme for the book.

To carry out our development more succinctly, consider the following terminology. We call the unconstrained model the *full model* because it is "full" of parameters, with the number of parameters in the full model frequently equaling the number of groups in the design. In the full model for the one-group case, we have one unknown parameter $\mu$, which is to be estimated on the basis of the data. The general method used to arrive at a second model is to place restrictions on the parameters of the first model. The restrictions are essentially our null hypothesis and serve to delete some of the parameters from the set used by the full model. We call the resultant constrained model simply the *restricted model.* In the one-group case, the restricted model does not require the estimation of any parameters. Although that is not usually the case in other designs, it is true that the restricted model always involves the estimation of fewer parameters than does the full model. Thus, we have the following models, least-squares estimates, and errors, in the one-group case:

| | **Model** | **Least-Squares Estimates** | **Errors** |
|---|---|---|---|
| Full: | $Y_i = \mu + \varepsilon_{i_\mathrm{F}}$ | $\hat{\mu} = \overline{Y}$ | $\sum e_{i_\mathrm{F}}^2 = \sum(Y_i - \overline{Y})^2$ |
| Restricted: | $Y_i = \mu_0 + \varepsilon_{i_\mathrm{R}}$ | No parameters estimated | $\sum e_{i_\mathrm{R}}^2 = \sum(Y_i - \mu_0)^2$ |

We use $E_\mathrm{F}$ to designate the sum of squared errors $\sum e_{i_\mathrm{F}}^2$ in the full model, and $E_\mathrm{R}$ to designate the analogous quantity $\sum e_{i_\mathrm{R}}^2$ for the restricted model.[3] Letting PIE stand for the proportional increase in error, we can express our verbal equation comparing the adequacy of the two models in algebraic form as

$$\mathrm{PIE} = \frac{E_\mathrm{R} - E_\mathrm{F}}{E_\mathrm{F}} \tag{19}$$

Substituting, we have

$$\begin{aligned}\mathrm{PIE} &= \frac{\sum e_{i_\mathrm{R}}^2 - \sum e_{i_\mathrm{F}}^2}{\sum e_{i_\mathrm{F}}^2} \\ &= \frac{\sum(Y_i - \mu_0)^2 - \sum(Y_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}\end{aligned}$$

and using Equation 17 to simplify the numerator, we obtain

$$\text{PIE} = \frac{n(\overline{Y} - \mu_0)^2}{\sum(Y_i - \overline{Y})^2} \tag{20}$$

Hopefully, the final way PIE is expressed looks at least vaguely familiar. One of the first hypothesis tests you likely encountered in your first statistics course was a one-sample $t$ test. Recall that the form of a one-sample $t$ test assessing the null hypothesis $H_0 : \mu = \mu_0$ looks at the deviation of a sample mean from the hypothesized value relative to the standard error of the mean

$$\begin{aligned} t &= \frac{\overline{Y} - \mu_0}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y} - \mu_0}{s/\sqrt{n}} \\ &= \frac{\sqrt{n}(\overline{Y} - \mu_0)}{\sqrt{\sum(Y_i - \overline{Y})^2/(n-1)}} \end{aligned} \tag{21}$$

where $\hat{\sigma}_{\overline{Y}}$ is the standard error of the mean (that is, the standard deviation of the sampling distribution of $\overline{Y}$) and $s$ is the square root of the unbiased sample variance. Note that if we were to square the form of the one-sample $t$ given on the right in Equation 21, we would have something very much like our PIE. In fact, all we would have to do to change PIE into $t^2$ is to divide the denominator[4] of the PIE by $(n - 1)$. (Note that we have said nothing about distributional assumptions; we are simply pointing out the similarity between how we would compute an intuitively reasonable statistic for comparing two models and the form of the test statistic for the one-sample $t$. We consider assumptions about the distribution of $Y$ scores shortly.)

We began our discussion of the model-comparison approach by noting that we want models that are simple yet adequate. You may wonder if we could not incorporate both of these aspects into a summary measure for comparing models. We must, in fact, do so. PIE simply compares the adequacy of the models (actually, in comparing errors of prediction, it does so by contrasting the inadequacy of the models) without regard to their complexity. To make PIE a more informative summary of the relative desirability of the models, we really want to take into account the simplicity of the models. We know in advance that our simpler, restricted model is necessarily less adequate than our full model (see Equation 17). Thus, intuitively, we would like our summary measure to indicate something such as, Is the loss in adequacy per additional unit of simplicity large? However, how could we assess the simplicity of a model?

The simplicity of a linear model is determined by the number of parameters: the fewer parameters, the simpler the model. As we illustrate momentarily, each parameter that we must estimate entails the loss of a degree of freedom. In fact, we define the degrees of freedom (*df*) resulting from using a particular equation as a model for an experiment as the number of independent observations in the study minus the number of independent parameters estimated. Thus, the *df* associated with a model can be used as our index of its simplicity. Given that, for a study having a fixed number of observations, the number of *df* associated with a model is inversely related to the number of parameters in the model, the *df* can be taken as a direct indicator of the model's simplicity: the more *df*, the simpler the model.

This allows us to construct a very useful summary measure for comparing models. The error of our more adequate model relative to its *df* gives us a basis for evaluating the size of the increase in error entailed by adopting a simpler model relative to the corresponding increase in *df*.

We can easily incorporate this consideration of the models' simplicity into our measure of the proportional increase in error.

Specifically, we need only divide the denominator and numerator of PIE in Equation 19 by the *df* of the model(s) involved in each. That is, in the denominator we divide the error of the full model ($E_F$) by the degrees of freedom of the full model ($df_F$), and in the numerator we divide the difference between the error of the restricted model and the error of the full model ($E_R - E_F$) by the difference in the degrees of freedom associated with the two models ($df_R - df_F$). This yields a revised measure, which we denote by $F$, of the relative adequacy yet simplicity of the two models:

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} \tag{22}$$

This simple comparison measure is in fact extremely useful and general. We can use it for carrying out all the hypothesis tests we need for the various special cases of the general linear model we will consider. All tests in ANOVA, analysis of covariance, bivariate regression, and multiple regression can be computed using this formula. The models being compared may differ widely from one situation to the next, but our method of comparing them can always be the same.

If there is no difference between the two models' descriptive accuracy except for the additional free parameter(s) in the full model, then the numerator (the increase in error per additional degree of freedom associated with using the simpler, restricted model) would be expected to be approximately the same as the denominator (the baseline indication of error per degree of freedom). Thus, values of $F$ near 1 would indicate no essential difference in the accuracy of the models, and the simpler model would be preferred on grounds of parsimony. However, if the increase in error associated with using the simpler model is larger than would be expected given the difference in parameters, then larger $F$ values result, and we tend to reject the simpler model as inadequate.

For the two models we are considering for a design involving only one group of subjects, we can determine the degrees of freedom to use in our general formula quite easily. In the full model, we are estimating just one parameter, $\mu$; thus, if we have $n$ independent observations in our sample, the degrees of freedom associated with the full model is $n - 1$. In the restricted model, we do not have to estimate any parameters in this particular case; thus, $df_R = n$. When we subtract $df_F$ from $df_R$, the number of subjects "drops out," and the difference is only the difference in the numbers of parameters estimated by the two models. Thus, for the one-group situation, we have

$$\begin{aligned} F &= \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} \\ &= \frac{n(\overline{Y} - \mu_0)^2/[n - (n - 1)]}{\sum(Y_i - \overline{Y})^2/(n - 1)} = t^2 \end{aligned} \tag{23}$$

To make this intuitively developed descriptive statistic useful for inferential purposes (i.e., hypothesis testing), we need only assume that the individual errors have certain characteristics. Specifically, if we assume the error terms $\varepsilon_i$ in our models are independently distributed as normal random variables with zero mean and variance $\sigma^2$, then it can be shown that the $F$ in our general formula does in fact follow a theoretical $F$ distribution with $df_R - df_F$ and $df_F$ degrees of freedom.

**TABLE 3.1**
HYPERACTIVE CHILDREN'S WISC-R SCORES

*Full-Model Analysis*

| *IQ Scores* $Y_i$ | *Prediction Equations* | *Parameter Term* $\hat{\mu}$ | *Error Scores* $e_{i_F} = Y_i - \hat{\mu}$ | *Squared Errors* $e_{i_F}^2$ |
|---|---|---|---|---|
| 96 | $= \hat{\mu} + e_1$ | 104 | $-8$ | 64 |
| 102 | $= \hat{\mu} + e_2$ | 104 | $-2$ | 4 |
| 104 | $= \hat{\mu} + e_3$ | 104 | 0 | 0 |
| 104 | $= \hat{\mu} + e_4$ | 104 | 0 | 0 |
| 108 | $= \hat{\mu} + e_5$ | 104 | $+4$ | 16 |
| 110 | $= \hat{\mu} + e_6$ | 104 | $+6$ | 36 |
| $\sum = 624$<br>$\bar{Y} = 104$ | | | $\sum = 0$ | $E_F = 120$ |

*Restricted-Model Analysis*

| *IQ Scores* $Y_i$ | *Prediction Equations* | *Parameter Term* $\mu_0$ | *Error Scores* $e_{i_R} = Y_i - \mu_0$ | *Squared Errors* $e_{i_R}^2$ |
|---|---|---|---|---|
| 96 | $= \mu_0 + e_1$ | 98 | $-2$ | 4 |
| 102 | $= \mu_0 + e_2$ | 98 | 4 | 16 |
| 104 | $= \mu_0 + e_3$ | 98 | 6 | 36 |
| 104 | $= \mu_0 + e_4$ | 98 | 6 | 36 |
| 108 | $= \mu_0 + e_5$ | 98 | 10 | 100 |
| 110 | $= \mu_0 + e_6$ | 98 | 12 | 144 |
| | | | | $E_R = 336$ |

$$F = \frac{(E_R - E_F/(df_R - df_F)}{E_F/df_F} = \frac{(366-120)/(6-5)}{120/5} = \frac{216}{24} = 9$$

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{\bar{Y} - \mu_0}{\sqrt{\dfrac{\sum(Y-\bar{Y})^2}{n-1}} \Big/ \sqrt{n}} = \frac{104 - 98}{\sqrt{\dfrac{120}{5}} \Big/ \sqrt{6}} = \frac{6}{\sqrt{\dfrac{24}{6}}} = 3$$

## Numerical Example

Assume that you work in the research office of a large school system. For the last several years, the mean score on the WISC-R, which is administered to all elementary school children in your district, has been holding fairly steady at about 98. A parent of a hyperactive child in one of your special education programs maintains that the hyperactive children in the district are actually brighter than this average. To investigate this assertion, you randomly select the files of six hyperactive children and examine their WISC-R scores. Table 3.1 shows these scores.

The unconstrained, or full, model does not make any a priori judgments about the mean IQ of hyperactive children. Rather, the estimate of $\mu$ is chosen so that $E_F = \Sigma e_{i_F}^2$ is minimized for this set of data. As we know, the sample mean, which here equals $624/6 = 104$, minimizes this sum of squared errors. Computing the deviations from this estimated population mean, we note that they sum to zero. This is, of course, always going to be the case because

$$\begin{aligned}\sum e_{i_F} &= \sum (Y_i - \bar{Y}) = \sum \left(Y - \sum Y/N\right) \\ &= \sum Y - \sum \left(\sum Y/N\right) = \sum Y - N\left(\sum Y/N\right) = 0\end{aligned}$$

We square each of these error scores and sum to obtain what we use as our index of the inadequacy of the model, that is, $E_F = 120$.

The degrees of freedom, which is the number of data values you would be free to choose once all parameter estimates have been specified, reflects the model's simplicity, as we indicated. For example, in the full model, once the sample mean is determined to be 104, you could choose five of the data values to be whatever you like, but the sixth must be the value that would bring the total to 624 so that the mean of the six scores will in fact be 104, that is, $Y_6 = 6(104) - \Sigma_{i=1}^{5} Y_i$. As indicated in Table 3.1, the *df* for our full model is 5—that is, the number of independent observations in the sample (6) minus the number of parameters estimated (1, which here is $\mu$). In general, the degrees of freedom associated with a model for a particular set of data is the total number of independent observations minus the number of parameters to be estimated in that model.

The analysis for the restricted model proceeds similarly. However, in this simplest case, there are no parameters to estimate, the average of the population having been hypothesized to be exactly 98. Thus, the error scores associated with this model can be computed directly by subtracting 98 from each score. When these error scores are squared and summed, we get a total error ($E_R = 336$) that is considerably larger than that associated with the full model ($E_F = 120$). Recall that the restricted model always has as great or greater summed errors than that associated with the full model. In fact, as shown (see Equations 17 and 20), the increase in error here depends simply on how far $\overline{Y}$ is from $\mu_0$, that is,

$$\begin{aligned} E_R - E_F &= n(\overline{Y} - \mu_0)^2 \\ &= 6(104 - 98)^2 = 6(6)^2 = 6(36) = 216 \\ &= 336 - 120 \end{aligned} \tag{24}$$

Finally, the degrees of freedom for the restricted model is simply equal to the number of observations—that is, 6—because no parameters had to be estimated.

Dividing our error summary measures by the corresponding degrees of freedom, as shown in our basic equation for the *F* near the bottom of Table 3.1, we obtain the values of the numerator and denominator of our test statistic. The value of 24 in the denominator is the squared error per degree of freedom for our full model (often referred to as *mean square error*). The value of 216 in the numerator is the increase in error per additional degree of freedom gained by adopting the restricted model. Computing their ratio, we get a value of 9 for *F*, which can be viewed, as we have indicated, at a descriptive level as an "adequacy yet simplicity" score. Its value here indicates that the additional error of the simpler restricted model per its additional degree of freedom is nine times larger than we would expect it to be on the basis of the error of the full model per degree of freedom. That is, the restricted model is considerably worse per extra degree of freedom in describing the data than is the full model relative to its degrees of freedom. Thus, intuitively it would seem that the restricted model should be rejected. We need, however, a statistical criterion for judging how large the *F* is.

To determine if the probability of obtaining an *F* this extreme is sufficiently small to justify rejecting the restricted model, we can consult the tabled values of the *F* distribution shown in Appendix Table A.2. To obtain a critical *F* value from the table, we consult the column corresponding to the degrees of freedom from the numerator of our test statistic—that is, $df_R - df_F$—and the main row of the table corresponding to the denominator degrees of freedom, that is, $df_F$. The third factor to be considered is the $\alpha$ level, that is, the probability of obtaining an *F* value larger than the tabled value, assuming that the restricted model is in fact correct. Critical *F* values are provided for six different $\alpha$ levels, namely .25, .10, .05, .025,

.01, and .001, on six adjacent rows of the table for each denominator *df*. When the observed $F$ value of 9 is compared against the tabled values of the $F$ distribution with numerator and denominator degrees of freedom of $df_{\mathrm{R}} - df_{\mathrm{F}} = 1$ and $df_{\mathrm{F}} = 5$, respectively, we find it exceeds the critical value of 6.61 for $\alpha = .05$. The conclusion would then be that there is significant reason to doubt that the population of hyperactive children has the same mean IQ as the other students in your district. The parent who brought the matter to your attention apparently was correct.

## Relationship of Models and Hypotheses

As may be clear, the two models being compared are the embodiments of two competing hypotheses. The full model corresponds to the alternative hypothesis, and the restricted model to the null hypothesis. In the full model and the alternative hypothesis, the population parameter is not constrained to equal any particular value. The restricted model is obtained from the full model by imposing the restriction on its parameters stated in the null hypothesis. As indicated below, restricting the $\mu$ in the full model to a particular value, $\mu_0$, such as 98, yields the restricted model:

| **Hypothesis** | **Model** | |
|---|---|---|
| $H_1 : \mu \neq \mu_0$ | Full: $Y_i = \mu + \varepsilon_i$ | (25) |
| $H_0 : \mu = \mu_0$ | Restricted: $Y_i = \mu_0 + \varepsilon_i$ | (26) |