Simple Models

Statistical Inferences about Parameter Values

4

In Chapter 2 we considered various alternatives for how we should measure aggregate error in the equation:

DATA = MODEL + ERROR

Although the sums of both the absolute errors and the squared errors seem to be reasonable alternatives, we decided in Chapter 3, for reasons of efficiency, practicality, and tradition, to define total error as the sum of the squared errors. In the simplest models, where we are estimating the single parameter β_0, the choice of the sum of squared errors as the definition of error implies that the best estimate is the sample mean (as we proved in Chapter 2).

In this chapter, we develop procedures for asking questions or testing hypotheses about simple models. Defining and answering interesting questions is the purpose of data analysis. We first consider the logic of answering questions about data for the case of the simplest models because it is easy to focus on the logic when the models are simple and because the logic generalizes easily to more complex models. The specific statistical test presented in this chapter is equivalent to the "one-sample *t*-test." We do not derive this test in terms of the *t* statistic; we prefer instead to construct this test using concepts and procedures that are identical to those required for the more complex models we will consider later.

The generic problem is that we have a batch of data for which we have calculated b_0, the mean, as an estimate of β_0. Our question is whether β_0 is equal to some specific value. For example, we might want to know whether the body temperatures of a group of patients administered a therapeutic drug differed from the normal body temperature of 37°C. We will let B_0 equal the value specified in our question. The statement:

$$\beta_0 = B_0$$

represents our *hypothesis* about the true value of β_0. Such statements are often called *null hypotheses*. The calculated value of b_0 will almost never exactly equal B_0, the hypothesized value of β_0. That is, the compact model:

MODEL C: $Y_i = B_0 + \varepsilon_i$

in which no parameters are estimated will almost always produce a larger error than the augmented model:

MODEL A: $Y_i = \beta_0 + \varepsilon_i$

in which β_0 is estimated by b_0, the mean of the batch of data. We will calculate PRE, the proportional reduction in error index developed in Chapter 1, to see how much better the predictions of Model A are than those of Model C. The original question then becomes not whether Model A is better than Model C, but whether Model A is better *enough* than Model C that we should reject the hypothesis that $\beta_0 = B_0$. Deciding what value of PRE is "better enough" is the essence of statistical inference and is the focus of this chapter.

To be less abstract, we will consider a detailed example. Suppose that 20 tickets were available for a lottery that had a single prize of $1000. How much would individuals be willing to pay for a 1 in 20 or 5% chance of winning the $1000 prize? The expected value of a ticket in this particular lottery would be $1000/20 tickets or $50. One might hypothesize that people would focus on the magnitude of the prize (i.e., the $1000 payoff) and thus be willing to pay more than the expected value of a ticket (i.e., $50). But one might also hypothesize that people would focus on the likelihood of losing whatever amount they paid and thus be generally willing to pay less than the $50 expected value of a ticket. Formally, we are comparing these two models:

MODEL A: Bid $= \beta_0 + \varepsilon_i$

MODEL C: Bid $= 50 + \varepsilon_i$

Suppose that 20 individuals participated in our hypothetical lottery by submitting the following bids to buy a ticket:

41	50	51	28	29
24	82	37	42	37
45	50	37	22	52
25	53	29	65	51

These data are displayed in Figure 4.1 in what is called a Stem and Leaf plot, where the left column of numbers indicates the left-most digit of each bid (the values of ten) and the numbers to the right of this column indicate the second digit of each bid (i.e., there were six bids in the twenties: 22, 24, 25, 28, 29, 29).

The average bid is $42.5, which is obviously less than the expected value of $50. However, we want to know whether these bids are really different from $50 or whether their mean is below $50 simply as a result of random variation in the data. In other words, even if the true state of the world were such that people generally are willing to pay the expected value of $50, it is unlikely that the average bid in our sample would equal $50 *exactly*. So, we need to determine whether the average bid of 42.5 that we obtained is different *enough* from 50 that we would be willing to conclude that, on average, people are willing to pay less than the expected value. In this example, β_0 represents the *true* typical amount that people would be willing to pay for a lottery ticket. We do not know, nor can we ever know, exactly what this *true* value is. The *hypothesized* value for β_0 is B_0, and in this example it equals 50,

FIGURE 4.1 Stem and Leaf plot for the 20 lottery bids

2	245899
3	777
4	125
5	001123
6	5
7	
8	2

Stem width = 10.00

the expected value of a ticket in our lottery. Note that \$50 was not estimated from our data. Rather, \$50 is an a priori hypothesized value: it is a specific value that was determined before looking at the data. The *estimated* value for β_0 is b_0, and in this case it equals 42.5, the mean or average of the 20 bids, because the mean minimizes the sum of squared errors. In other words, for the compact model (which represents the null hypothesis), the prediction is given by:

$$\text{MODEL C: } \hat{Y}_i = 50$$

and for the augmented model (in which β_0 is estimated from the data), the prediction is given by:

$$\text{MODEL A: } \hat{Y}_i = 42.5$$

So our question is whether the predictions of Model A are better enough to infer that Model C is unreasonable.

To answer our question, we want to calculate PRE. To do so, we first need to calculate the error for each model. The necessary calculations are displayed in Figure 4.2. For the compact model, $\hat{Y}_{iC} = 50$, so the sum of squared errors from the compact model, SSE(C), is given as:

$$\text{SSE(C)} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{iC})^2 = \sum_{i=1}^{20}(Y_i - 50)^2 = 5392$$

The squared error using the compact model for each bidder is listed in the third column of Figure 4.2, along with the sum of 5392 for SSE(C). For the augmented model, $\hat{Y}_{iA} = 42.5$, so:

$$\text{SSE(A)} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{iA})^2 = \sum_{i=1}^{20}(Y_i - 42.5)^2 = 4267$$

The squared error using the augmented model for each bidder is listed in the fourth column, along with its sum of 4267 for SSE(A). Then the proportional reduction in error using Model A instead of Model C is given by:

$$\frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{5392 - 4267}{5392} = .209$$

That is, Model A using b_0, the *estimated* value of β_0, has 20.9% less error than Model C using B_0, the *hypothesized* value of β_0. Later in this chapter, we will determine whether 20.9% less error is enough to warrant rejecting Model C (\$50) in favor of Model A.

We note in passing that for Model A one observation (bidder 7) is responsible for a substantial proportion of the total SSE. Although the presentation of formal procedures for investigating outliers must wait until Chapter 13, large errors associated with a few observations should make us suspect the presence of outliers. Remember that SSE and its associated estimators, such as the mean, are not resistant to outliers.

FIGURE 4.2 Lottery bids and error calculations for 20 bidders in a hypothetical lottery

		Squared errors	
	Bid	*Compact*	*Augmented*
Bidder number	Y_i	$(Y_i - B_0)^2$	$(Y_i - b_0)^2$
1	41	81	2.25
2	50	0	56.25
3	51	1	72.25
4	28	484	210.25
5	29	441	182.25
6	24	676	342.25
7	82	1024	1560.25
8	37	169	30.25
9	42	64	0.25
10	37	169	30.25
11	45	25	6.25
12	50	0	56.25
13	37	169	30.25
14	22	784	420.25
15	52	4	90.25
16	25	625	306.25
17	53	9	110.25
18	29	441	182.25
19	65	225	506.25
20	51	1	72.25
Sum	850.00	5392.00	4267.00
Mean	42.50		

DECOMPOSITION OF SSE

At this point it is useful to introduce a table that summarizes our analysis so far. The sum of squares reduced (SSR) is defined as:

$$\text{SSR} = \text{SSE(C)} - \text{SSE(A)}$$

and represents the amount of error that is reduced by using Model A instead of Model C. Then it is obvious that:

$$\text{SSE(C)} = \text{SSR} + \text{SSE(A)}$$

In other words, the original error SSE(C) can be decomposed into two components: (a) the reduction in error due to Model A (i.e., SSR) and (b) the error remaining from Model A (i.e., SSE(A)). It is common to summarize the results of an analysis in a table having separate rows for SSR, SSE(A), and SSE(C). Figure 4.3 provides the generic layout for such tables, which are referred to as analysis of variance or ANOVA tables because they analyze (i.e., separate or partition) the original variance or error into component parts. Figure 4.4 presents the ANOVA summary table for our example. Note that the SS (sum of squares) for the total line, which represents SSE(C), is indeed the sum of SSR and

FIGURE 4.3 Generic ANOVA layout

Source	*SS*	*PRE*
Reduction using Model A	SSR	SSR/SSE(C)
Error for Model A	SSE(A)	
Total	SSE(C)	

FIGURE 4.4 ANOVA summary table for lottery example

Source	*SS*	*PRE*
Reduction (using β_0)	1125	.209
Error (using β_0)	4267	
Total (error using B_0)	5392	

SSE(A); for our example, 5392 = 1125 + 4267. PRE is readily obtained from the SS column using the formula:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(C)}}$$

We will use these tables, which give the decomposition of the sums of squares, as the basic summary for all our statistical tests. Later, we will add several other useful columns to such tables.

SSR is easily understood and often easily calculated as the difference between SSE(C) and SSE(A). However, there is another representation for SSR, which provides additional insight for the comparison of Models C and A. It can be shown (we essentially did it in Chapter 2 for the case of the simple model; the more general proof does not provide useful insights so we omit it) that:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_{iC} - \hat{Y}_{iA})^2 \tag{4.1}$$

where $\hat{Y}_{iC}$ and $\hat{Y}_{iA}$ are, respectively, the predictions for the *i*th observation using Model C and Model A. This formula will be useful later for calculating certain SSRs that are not automatically provided by typical computer programs. More important are the insights it provides. For a fixed SSE(C), the larger SSR is, the larger PRE is, and the larger the improvement provided by using Model A instead of Model C. This formula shows that SSR is small when Models C and A generate similar predictions for each observation. In the extreme case when Models C and A are identical (i.e., they produce the same predictions), then SSR = 0 and PRE = 0. Conversely, SSR will be large to the extent that Models C and A generate *different* predictions. Thus, SSR is a direct measure of the difference between Models C and A and PRE = SSR/SSE(C) is a proportional measure of that difference.

Equation 4.1 is useful for calculating the SSR for the simple models considered in this chapter. Although we generally avoid multiple computational formulas, we present this one because many computer programs do not conveniently provide the necessary information for computing PRE in our terms. We will illustrate the use of Equation 4.1 by computing the SSR for the lottery example. The value predicted by MODEL A is $\hat{Y}_{iA} = \bar{Y} = 42.5$ and the value predicted by Model C is the hypothesized value $\hat{Y}_{iC} = B_0 = 50$. So, according to Equation 4.1:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_{iC} - \hat{Y}_{iA})^2 = \sum_{i=1}^{20} (50 - 42.5)^2 = \sum_{i=1}^{20} (7.5)^2 = 20(56.25) = 1125$$

That is, SSR equals the constant $7.5^2 = 56.25$ summed 20 times. Thus, SSR = 20(56.25) = 1125, which is the same value that we obtained by calculating SSR directly as SSE(C) – SSE(A) in Figure 4.3. Thus, for simple models, by comparing a Model A that estimates one parameter with a Model C that estimates none, the following formula is often handy:

$$\text{SSR} = \sum_{i=1}^{n} (B_0 - \bar{Y})^2 = n(B_0 - \bar{Y})^2 \tag{4.2}$$

We will have many occasions to use Equation 4.2.

SAMPLING DISTRIBUTION OF PRE

It might seem that a difference in parameter estimates of 50 – 42.5 = 7.5 and PRE = 20.9% are "large enough" to infer that Model C is unreasonable relative to Model A. Unfortunately, such a conclusion may not be warranted statistically, and it is important to understand why. To gain this understanding, we need to focus on the error term ε_i that is included in the full statement of Model C:

MODEL C: $Y_i = B_0 + \varepsilon_i$

As noted before, this model says that were it not for random perturbations represented by ε_i all the Y_i values would equal B_0 exactly. In Chapter 3 we made the assumption that the ε_i values are all sampled randomly and independently from a normal distribution with mean 0 and variance σ^2. We also saw in Chapter 3 that the exact value for the mean calculated from a sample of size n would depend on the particular sample of errors. Sometimes the calculated mean would be above the true value B_0, and sometimes it would be below. That is, there would be a sampling distribution for the mean. If Model C were correct, sometimes the sample mean would be somewhat higher than 50 and other times it would be somewhat lower, but it would seldom equal 50 exactly. For example, we would most likely have obtained a different mean if the bids had been gathered before lunch or the day before or a day later, because the pattern of random perturbations would have been different.

Similar to the sampling distribution for the mean, there is also a sampling distribution for PRE. Just as b_0, calculated from the data, is the estimate of the unknown true parameter β_0, so too PRE, calculated from the data, is the estimate of the unknown true proportional reduction in error η^2. For the moment, let us assume that Model C is correct (i.e., that $\beta_0 = B_0$) and consider what the sampling distribution for PRE would be. In other words, we begin by assuming that the null hypothesis is true. In these terms, $\eta^2 = 0$ is equivalent to Model C being correct and to Model A making absolutely no improvement relative to Model C. We saw in Chapter 3 that b_0 has a sampling distribution, so even if Model C is correct we would not expect our estimate b_0 to equal B_0 exactly. But we know that b_0 produces the smallest possible sum of squared errors, so SSE(A), using b_0, must always be at least a little smaller than SSE(C), using B_0. For example, in the lottery data the mean will seldom equal 50 exactly, even if the true parameter value were 50, and thus the SSE calculated using the sample mean will always be a little less than SSE(C) (see the proof in Chapter 2). Hence, even if the true proportional reduction in error $\eta^2 = 0$ (as it must when Model C is correct), the calculated PRE will always be at least a little greater than zero (never less than zero). PRE is therefore a biased estimator of η^2 because PRE will always overestimate the true value of η^2. We will return to this issue of the bias in PRE later. For now, the important point is that we should not expect the calculated PRE to equal zero even when Model A makes no improvement on Model C. Thus, we cannot base our decision about the validity of Model C simply on whether or not PRE = 0.

If we cannot use PRE ≠ 0 as a criterion for rejecting Model C, then we need to consider the sampling distribution of PRE to determine whether the calculated value of PRE is a *likely* value, *assuming that Model C is correct*. If the calculated value of PRE is a likely value, then we ought not to reject Model C and its equivalent null hypothesis that $\beta_0 = B_0$. On the other hand, if the calculated value of PRE is an unlikely value when Model C is assumed to be true, then we ought to reject Model C in favor of Model A. In terms of our example, we need to consider the sampling distribution of PRE to determine for this case whether PRE = .209 is a likely value, assuming that Model C is correct (i.e., that $\beta_0 = 50$). If it is a likely value, then we ought not to reject Model C and its equivalent hypothesis that $\beta_0 = B_0 = 50$; there would be no evidence that the lottery bids are different from what would be expected if the null hypothesis were true. If PRE = .209 is an unexpected value, then we ought to reject Model C and its hypothesis in favor of Model A and its estimate that $\beta_0 = b_0 = 42.5$; in other words, we would conclude that the lottery bids were significantly lower than the expected value of $50 for a ticket.

We could develop the sampling distribution for PRE in this example using the same simulation strategy we used in the previous chapter. That is, we could put error tickets of the appropriate size into a bag and then do many simulation rounds, in each of which we would randomly select 20 error tickets, add 50 to each, and calculate PRE for that sample. The only problem with this strategy is that we do not know what size error tickets to place in the bag to be sampled. In other words, we do not know the variance σ^2 of the normal distribution of errors. However, as was noted in the previous chapter, the mean squared error provides an estimate of σ^2. In particular, for Model A with one parameter the estimate is:

$$s^2 = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{4267}{19} = 224.58$$

We could therefore conduct the simulation by sampling error values from a normal distribution with mean 0 and variance 224.58. The 100 error tickets in Figure 3.1 were sampled from such a normal distribution, so they could be used as the bag of error tickets for the simulation.

As an example of a simulation round, suppose that the following 20 error tickets were drawn from Figure 3.1:

17	31	–2	–6	–17
–4	28	29	17	1
–12	–6	3	–25	2
24	15	–20	4	–3

These error terms when added to the value of $B_0 = 50$ of Model C yield the following 20 scores:

67	81	48	44	33
46	78	79	67	51
38	44	53	25	52
74	65	30	54	47

The mean of the resulting 20 scores is 53.8 and SSE(C), using 50 as the model prediction for all the observations, and SSE(A), using 53.8 as the model prediction, are easily calculated to be 5554 and 5265.2, respectively. Thus:

$$\text{PRE} = \frac{5554 - 5265.2}{5554} = .052$$

Then a new simulation round with a new sample of error values would produce a different mean and PRE. These simulation rounds could conceptually be repeated until there were enough PRE values to make a sampling distribution for PRE.

Alas, the simulation strategy outlined above for generating the sampling distribution of PRE will not work because $s^2 = 224.58$ is only an *estimate* of the true variance of the errors σ^2. Just as it is unlikely that $\bar{Y} = b_0$ (the calculated mean for a sample) will equal B_0 exactly, it is also unlikely that the calculated variance s^2 will equal σ^2 exactly. In other words, we are uncertain about the exact size of the error tickets that should be placed in the bag.

We could conduct a more complex sampling simulation to account for our uncertainty about the size of the error tickets. However, it would be tedious if we had to do a new simulation round for each data analysis. Fortunately, this is not necessary because mathematical statisticians have specified the sampling distribution for PRE based on the assumptions we made in Chapter 3 about the behavior of the errors ε_i. Even though we will not actually do simulations for generating a sampling distribution, it is important to remember that the mathematical formula for that distribution is derived from the assumption that the error values are randomly sampled from a distribution with mean 0 and variance σ^2 and that the sampling could be represented by drawing error tickets from a bag.

Figure 4.5 provides a tabular description of the sampling distribution of PRE for the particular case of samples of size 20, again assuming the validity of Model C. That is, if Model C were correct (in our case, $\beta_0 = 50$), and if we compared Model C ($\hat{Y} = 50$) with Model A ($\hat{Y} = b_0 = \bar{Y}$) from samples, then Figure 4.5 presents the proportional frequency and cumulative proportion for each range of PRE. The proportional frequencies are plotted in Figure 4.6. As we would expect, the sampling distribution in Figure 4.6 shows that values of PRE near zero are the most likely. It also shows that values of PRE greater than .2 are infrequent.

The cumulative proportions are generally more useful than the proportions for individual ranges. The cumulative proportion is simply the total proportion for the range of PREs from zero to the value of interest. For example, to find the cumulative proportion for the range from 0 to .03, we simply add the proportions for the three component ranges: 0 to .01, .01 to .02, and .02 to .03. For this range, .334 + .125 + .088 = .547. That is, 54.7% of the simulated PRE values are less than or equal to .03. The cumulative

FIGURE 4.5 Tabular description of the sampling distribution of PRE for testing the simple model with 20 observations

PRE range	*Proportion*	*Cumulative proportion*
.00–.01	.334	.334
.01–.02	.125	.459
.02–.03	.088	.547
.03–.04	.068	.615
.04–.05	.055	.670
.05–.06	.045	.716
.06–.07	.038	.754
.07–.08	.032	.786
.08–.09	.028	.814
.09–.10	.024	.838
.10–.11	.021	.858
.11–.12	.018	.876
.12–.13	.016	.892
.13–.14	.014	.905
.14–.15	.012	.917
.15–.16	.010	.928
.16–.17	.009	.937
.17–.18	.008	.945
.18–.19	.007	.952
.19–.20	.006	.958
.20–.21	.005	.963
.21–.22	.005	.968
.22–.23	.004	.972
.23–.24	.004	.976
.24–.25	.003	.979
.25–.26	.003	.982
.26–.27	.002	.984
.27–.28	.002	.986
.28–.29	.002	.988
.29–.30	.002	.990
.30–.31	.001	.991
.31–.32	.001	.993
.32–.33	.001	.994
.33–.34	.001	.995
.34–.35	.001	.995
.35–.36	.001	.996
.36–.37	.001	.997
.37–.38	.001	.997
.38–.39	.000	.998
.39–.40	.000	.998

FIGURE 4.6 Sampling distribution of PRE for testing the simple model with 20 observations

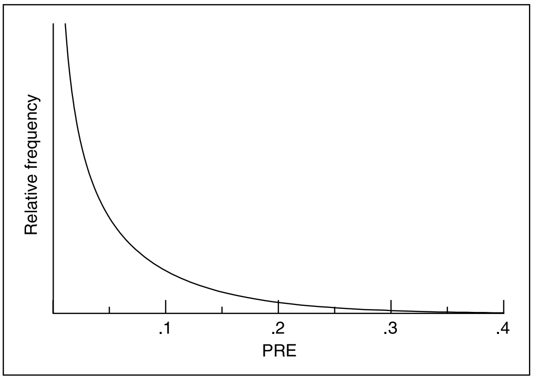
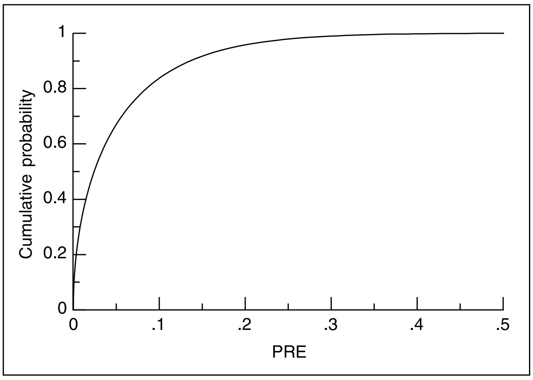


FIGURE 4.7 Cumulative proportions for *n* – PA = 19 (PA = number of parameters for Model A)



proportions are displayed in the last column of Figure 4.5 and graphed in Figure 4.7. We can see, for example, from both the table and the graph of the cumulative proportions that PRE is less than .1 about 84% of the time.

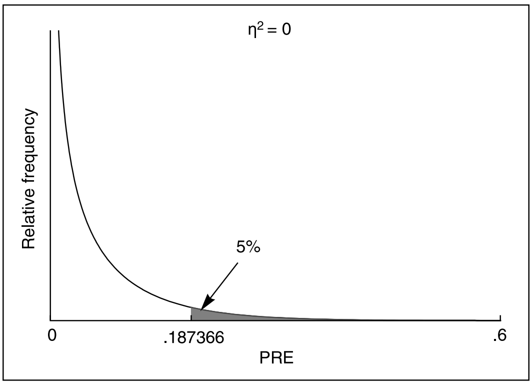
We can now ask whether a value for PRE of .209 is likely if Model C is correct. From the table in Figure 4.5 we see that approximately 96% of the time PRE would be less than .209. Or, in other words, a PRE as large as .209 would be obtained only about

4% of the time if Model C were correct. We can finally answer our question. It is unlikely (less than 5% chance) that we would have obtained a PRE this large had Model C been correct. We can therefore reasonably reject Model C in favor of Model A. This is equivalent to rejecting the null hypothesis that $\beta_0 = B_0 = 50$. Substantively, the data indicate that participants were willing to pay less than the expected value of the lottery tickets.

CRITICAL VALUES

From the mathematical equations describing the sampling distribution for PRE, we can determine for our example data that if Model C were correct (and thus $\eta^2 = 0$) we would expect 95% of the simulated values of PRE to be below the precise value of .187. This sampling distribution is plotted in Figure 4.8. In the social sciences it is customary to consider a value of PRE to be surprising if it occurs by chance less than 5% of the time when the null hypothesis is true. Thus, .187 is the *critical value* for this example; any value of PRE > .187 causes us to reject Model C. Using the equations, for any number of observations we can calculate the value of PRE for which we would expect 95% of the simulated PRE values to be below if Model C were correct. Figure 4.9 gives the 95% (and 99%) critical values for selected numbers of observations.

FIGURE 4.8 Distribution of PRE for n – PA = 19, assuming that $\eta^2 = 0$ (PA = number of parameters for Model A)



STATISTIC *F*

Figures of the critical values for PRE are rare in statistics books. Much more common, for largely historical reasons, are tables of F, a statistic closely related to PRE. As we shall see below, F is a simple function of PRE, so if we know PRE, the number of observations, and the number of parameters in Models C and A, then we also know F and vice versa. By re-expressing PRE, F also provides additional insights about the proportion of error reduced. We therefore turn to the motivation for calculating F and then consider its sampling distribution.

FIGURE 4.9 Critical values (95% and 99%) for PRE and F for testing models that differ by one parameter (PA = number of parameters for Model A)

	95%		*99%*	
n – PA	*PRE*	*F*	*PRE*	*F*
1	.994	161.45	1.000	4052.18
2	.903	18.51	.980	98.50
3	.771	10.13	.919	34.12
4	.658	7.71	.841	21.20
5	.569	6.61	.765	16.26
6	.499	5.99	.696	13.75
7	.444	5.59	.636	12.25
8	.399	5.32	.585	11.26
9	.362	5.12	.540	10.56
10	.332	4.97	.501	10.04
11	.306	4.84	.467	9.65
12	.283	4.75	.437	9.33
13	.264	4.67	.411	9.07
14	.247	4.60	.388	8.86
15	.232	4.54	.367	8.68
16	.219	4.49	.348	8.53
17	.208	4.45	.331	8.40
18	.197	4.41	.315	8.29
19	.187	4.38	.301	8.19
20	.179	4.35	.288	8.10
22	.164	4.30	.265	7.95
24	.151	4.26	.246	7.82
26	.140	4.23	.229	7.72
28	.130	4.20	.214	7.64
30	.122	4.17	.201	7.56
35	.105	4.12	.175	7.42
40	.093	4.09	.155	7.31
45	.083	4.06	.138	7.23
50	.075	4.03	.125	7.17
55	.068	4.02	.115	7.12
60	.063	4.00	.106	7.08
80	.047	3.96	.080	6.96
100	.038	3.94	.065	6.90
150	.025	3.90	.043	6.81
200	.019	3.89	.033	6.76
500	.008	3.86	.013	6.69
∞		3.84		6.63

The two reasons for calculating *F* are (a) to examine the proportional reduction in error *per additional parameter* added to the model and (b) to compare the proportion of error that was reduced (PRE) with the proportion of error that remains (1 – PRE). In the context of the simple models that we are considering in this chapter, PRE is obtained by the addition of only one parameter. But later we will want to consider the improvement produced by models that add more than one parameter. To avoid having to present different formulas as the models become more complex, we will present the general definition of *F* here. The key idea is that a given PRE, let us say .35, is more impressive when obtained by the addition of a single parameter than when it is obtained by the addition of several parameters. So, we want to consider *PRE per parameter*. That is, we divide PRE by the number of additional parameters used in Model A that are not used in Model C. We will let PA and PC represent the number of parameters for Model A and Model C, respectively. Then, the number of additional parameters is simply PA – PC. Hence, F is based on the quantity:

$$\frac{\text{PRE}}{\text{PA} - \text{PC}}$$

which is simply the proportional reduction in error per additional parameter. For the simple models of this chapter, there are no parameters to be estimated for Model C and only one for Model A, so PC = 0, PA = 1, and PA – PC = 1.

Similarly, we need to consider the remaining proportion of the error, 1 – PRE, in terms of the number of additional parameters that *could* be added to reduce it. As noted in Chapter 1, the most parameters we can have is one for each observation Y_i. If there are n observations and we have already used PA parameters in Model A, then at most we could add n – PA parameters to some more complex model. So:

$$\frac{1 - \text{PRE}}{n - \text{PA}}$$

is the proportion of remaining error per parameter that *could* be added to the model. In other words, this is the average remaining error per additional parameter. If we added a parameter to the model at random, even a parameter that was not really useful, we would expect at least some reduction in error. The proportion of remaining error per parameter or the average remaining error tells us the value of PRE to expect for a worthless parameter. If the parameter or parameters added to the model are genuinely useful, then the PRE per parameter that we actually obtain ought to be substantially larger than the expected PRE per parameter for a useless, randomly selected parameter. An easy way to compare PRE per parameter obtained with the expected PRE per parameter is to compute the ratio of the two quantities; this gives the definition of F as:

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1-\text{PRE})/(n-\text{PA})} \tag{4.3}$$

We can think of the numerator of F as indicating the average proportional reduction in error per parameter added, and the denominator as the average proportional reduction in error that could be obtained by adding all possible remaining parameters. For Model A to be significantly better than Model C, we want the average error reduction for the parameters added to be much greater than the average error reduction we could get by adding the remainder of the possible parameters. Hence, if F is about 1, then we are doing no better than we could expect on average, so values of F near 1 suggest that we should not reject the simpler Model C. Values of F much larger than 1 imply that the average PRE per parameter added in Model A is much greater than the average that could be obtained by adding still more parameters. In that case, we would want to reject Model C (and its implicit hypothesis) in favor of Model A. For the example of Figure 4.2 where PRE = .209 and $n = 20$:

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} = \frac{.209/(1 - 0)}{.791/(20 - 1)} = \frac{.209}{.0416} = 5.02$$

In other words, we obtained a 20.9% reduction in error per parameter by using Model A, and the further reduction we could get by adding all the additional parameters is only 4.16% per parameter. Their ratio of 5.02 suggests that adding to Model C the one specific additional parameter β_0, which is estimated by the mean, yields a substantially better (about five times better) PRE than we could expect by randomly adding a parameter from the remaining ones. In other words, the increased complexity of Model A is probably worth it. But again we need to consider the sampling distribution of F to determine whether a value of 5.02 is indeed "surprising."

Again, mathematical equations exist for describing the sampling distribution of F, given the assumptions about error discussed in Chapter 3. If we assume that the errors ε_i are independently, identically, and normally distributed, then F has what is known as an *F distribution*. The 95% and 99% critical values for F for testing simple models are listed in Figure 4.9 next to their corresponding values of PRE. F and PRE are redundant in the sense that one exceeds its critical value if and only if the other one exceeds its critical value. For the example, PRE = .209 exceeds its critical value of .187, and so necessarily the corresponding $F = 5.02$ exceeds its critical value of 4.38. Thus, either PRE or F leads us to reject Model C and its implicit hypothesis that $\beta_0 = B_0 = 50$. Note that for most reasonable numbers of observations the 95% critical value for F is about 4. If we ignore the fractional part of F, then a useful rule of thumb that reduces the need to consult statistical tables frequently is to reject Model C in favor of Model A whenever F is greater than 5. If F is between 4 and 5, then you will probably have to look it up in the table, and if it is below 4, then there is no hope unless the number of observations is extremely large. Critical values of PRE and F for testing the more complex Model A that differs from Model C by more than one parameter are listed in the Appendix as a function of PA – PC, often called the "numerator degrees of freedom" because that term appears in the numerator of the formula for F (Equation 4.3), and n – PA, often called the "denominator degrees of freedom" because it appears in the denominator of the formula for F.

It is useful to add the degrees of freedom (df) and F to the basic summary table we began earlier. Figure 4.10 presents such a table for our example. It is our policy to avoid multiple computational formulas for the same quantity and instead to present only one conceptual formula. However, we must break that policy for F in this instance because F is traditionally calculated by an equivalent but different formula based on Figure 4.10. Figures constructed using the alternative formula for F are ubiquitous, so the reader has no choice but to learn this alternative in addition to the conceptual formula for F presented above. The alternative formula for F is:

$$F = \frac{\text{SSR}/(\text{PA} - \text{PC})}{\text{SSE(A)}/(n - \text{PA})} = \frac{\text{MSR}}{\text{MSE}}$$

For our example, this yields:

$$F = \frac{\text{SSR}/(\text{PA} - \text{PC})}{\text{SSE(A)}/(n - \text{PA})} = \frac{1125/(1-0)}{4267/(20-1)} = \frac{1125}{224.58} = 5.01$$

This agrees with our previous calculation except for a small rounding error. MSR represents the *mean squares reduced*, and MSE represents the *mean square error*. To facilitate this calculation, we usually add an "MS" column to the summary table. The final column, labeled "p," gives the probability of obtaining a PRE and F that is that large or larger if $\eta^2 = 0$. In this case, PRE and F exceed the 95% critical values so the probability of getting a PRE or F that is that large or larger if Model C were correct is less than .05. With the additional columns, Figure 4.11 provides a detailed summary of our analysis of the lottery bids.

FIGURE 4.10 Analysis of variance summary table: decomposition of sums of squares

Source	*SS*	*df*	*MS*	*F*	*PRE*	*p*
Reduce, Model A	SSR	PA – PC	$\text{MSR} = \frac{\text{SSR}}{\text{PA} - \text{PC}}$	$\frac{\text{MSR}}{\text{MSE}}$	$\frac{\text{SSR}}{\text{SSE(C)}}$	
Error for Model A	SSE(A)	n – PA	$\text{MSE} = \frac{\text{SSE (A)}}{n - \text{PA}}$			
Total	SSE(C)	n – PC				

FIGURE 4.11 ANOVA summary table for lottery example

Source	*SS*	*df*	*MS*	*F*	*PRE*	*p*
Reduce, Model A	1125	1	1125.00	5.01	.209	<.05
Error for Model A	4267	19	224.58			
Total	5392	20				

STATISTICAL DECISIONS

We have now defined the essence of statistical inference: if PRE and F exceed their respective critical values then the simpler Model C is rejected in favor of the more complex Model A. We now have a rule for resolving the inherent tension in data analysis between reducing the error as much as possible and keeping the model of the data as parsimonious as possible. It is important to recognize, however, that statistical inference is probabilistic and therefore not infallible. That is, if Model C is actually the correct model, then 5% of the time we will obtain values of PRE and F that exceed their 95% critical values. Our rule is to reject Model C if those statistics exceed their 95% critical values, so in such instances we will have made a mistake in rejecting Model C. There is no way to avoid making occasional mistakes of that type. By adopting a 95% critical value, we are implicitly accepting that for those data for which Model C is correct we are willing to risk a 5% chance of incorrectly rejecting it. Mistakes of this type are known as *Type I* errors. The choice of 5% as an acceptable rate of Type I errors is inherently arbitrary. If we want to be more cautious we could choose a rate of 1%, or if we are willing to risk more Type I errors then we might choose a rate of 10%.

We also can commit a *Type II* error. A Type II error occurs when Model C is incorrect but the obtained values of PRE and F still do not exceed their critical values. Thus, a Type II error occurs when we fail to reject Model C when we should. That is, Model C is incorrect and Model A is significantly better, but we are unlucky in terms of the error tickets drawn and miss seeing the difference. Figure 4.12 summarizes the statistical decision that confronts us and defines the ways in which both the right and wrong decisions can be made. Statistical inference can be viewed as a game with Nature. Nature determines whether Model C is correct or incorrect. The goal of the data analyst is to "guess" which is the case. The data analyst uses the data to make an informed guess. Specifically, if PRE and F exceed their critical values, then the decision is to "reject Model C"; otherwise the decision is "do not reject Model C." If Nature has determined that Model C is correct, then, using a 95% critical value, we will decide correctly 95% of the time and incorrectly (i.e., make a Type I error) 5% of the time. The chance of making a Type I error is often labeled α and referred to as the *significance level.*

On the other hand, if Nature has determined that Model C is incorrect, then we will decide correctly the proportion of times that PRE and F exceed their critical values, and we will decide incorrectly (i.e., make a Type II error) the proportion of times that PRE and F fall below their critical values. The chance of making a Type II error is, unfortunately, often labeled β, which should not be confused with β_0, β_1, etc., which we use to represent parameters in a model. The proportion of times the correct decision is made when Nature has determined that Model C is incorrect, $1 - \beta$, is often referred to as the *power* of a statistical test.

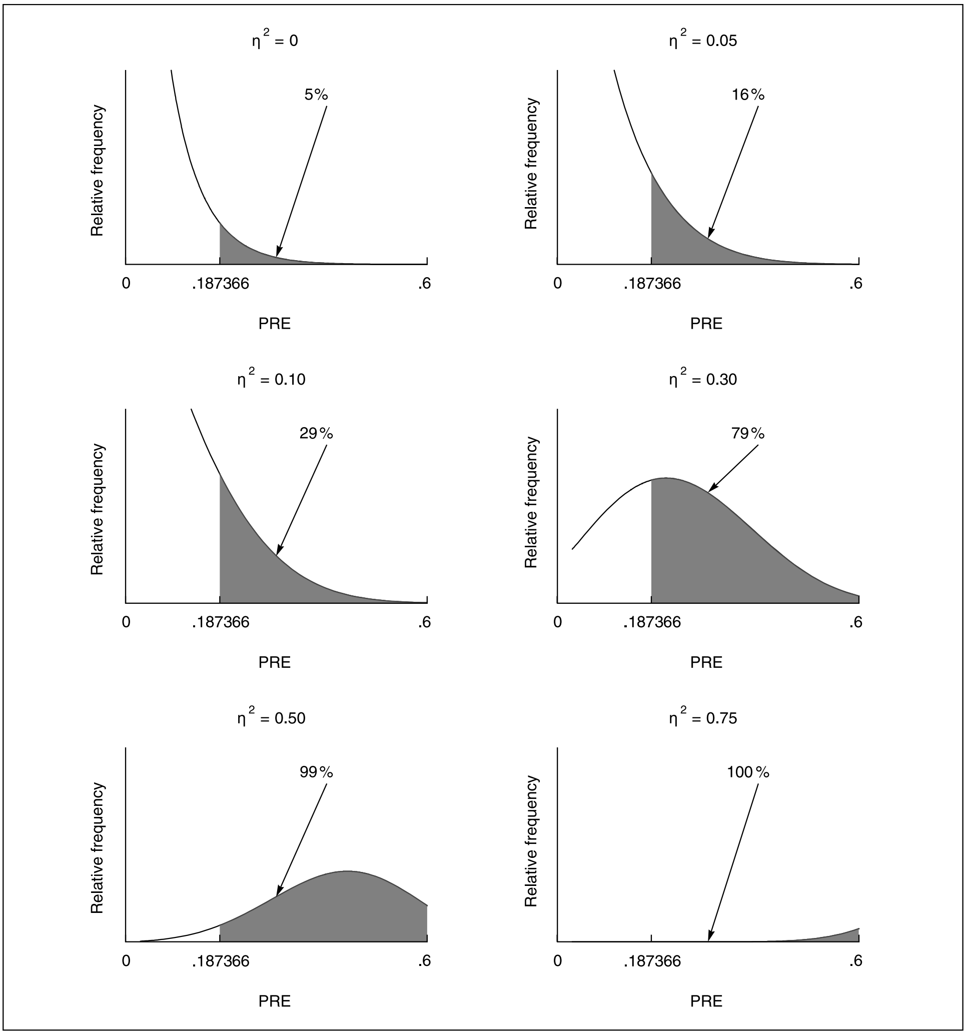
FIGURE 4.12 The statistical decision and the two types of errors

	True state of Nature	
Statistical decision	*Model C correct*	*Model C incorrect*
"Reject Model C"	Type I error	Correct decision
"Do not reject Model C"	Correct decision	Type II error

ESTIMATING STATISTICAL POWER

To determine the chances of a Type II error or to determine the power, we need to know the sampling distribution for PRE and F, *assuming that Model C is incorrect.* We cannot determine such a sampling distribution in general because to say that Model C is incorrect is to say only that the true proportional reduction in error η^2 is greater than zero. However, using the equations provided by mathematical statisticians, we can easily derive the sampling distributions for the specific values of η^2 we might want to consider. For our example in which PA – PC = 1 and n – PA = 19, Figure 4.13 displays plots of the sampling distribution for PRE assuming progressively greater true values of PRE, that is, η^2. Note that if $\eta^2 = .1$, the probability of obtaining a PRE greater than the critical value and thus rejecting Model C is 29% as compared with only 5% if Model C is correct and $\eta^2 = 0$. Figure 4.14 displays the cumulative probability distributions for other

FIGURE 4.13 Distributions of PRE for PA – PC = 1 and *n* – PA = 19, assuming various values for η^2



observed values of PRE if we were to assume that η^2 were equal to 0, .05, .1, .3, .5, and .75, but again only for the very particular conditions of our example: PA – PC = 1 and *n* – PA = 19. Different sampling distributions would be obtained for other combinations of PA – PC and *n* – PA. The column for $\eta^2 = 0$ corresponds exactly to the cumulative probability distribution of Figure 4.5. Each entry in Figure 4.14 is the probability that PRE calculated from the data would be less than the value of PRE specified for that row. For example, the value of .08 in the row for PRE = .30 and the column for η^2 = .50 means that if η^2 (i.e., the true PRE) were really .5, then the probability of obtaining a PRE (calculated from the data) of .3 or lower is 8%.

We can use the cumulative probability distributions of PRE for different values of η^2 to perform "what if" analyses. For example, we can ask, "*what* would the chances of making a Type II error be *if* η^2 = .3?" To answer this question, we first must decide what chance of a Type I error we are willing to risk. If we adopt the customary value

FIGURE 4.14 Cumulative sampling distributions of PRE for various η^2 when PA – PC = 1 and n – PA = 19

	True PRE, η^2					
PRE	*0*	*.05*	*.1*	*.3*	*.5*	*.75*
.00	.00	.00	.00	.00	.00	.00
.05	.67	.47	.31	.03	.00	.00
.10	.84	.65	.49	.08	.00	.00
.15	.92	.78	.63	.16	.01	.00
.20	.96	.86	.74	.25	.02	.00
.25	.98	.92	.83	.37	.04	.00
.30	.99	.95	.89	.49	.08	.00
.35	1.00	.97	.94	.61	.14	.00
.40	1.00	.99	.96	.72	.23	.00
.45	1.00	.99	.98	.81	.35	.00
.50	1.00	1.00	.99	.88	.48	.00
.55	1.00	1.00	1.00	.93	.62	.01
.60	1.00	1.00	1.00	.97	.75	.04
.65	1.00	1.00	1.00	.99	.86	.11
.70	1.00	1.00	1.00	1.00	.94	.25
.75	1.00	1.00	1.00	1.00	.98	.47
.80	1.00	1.00	1.00	1.00	.99	.73
.85	1.00	1.00	1.00	1.00	1.00	.92
.90	1.00	1.00	1.00	1.00	1.00	.99
.95	1.00	1.00	1.00	1.00	1.00	1.00

in the social sciences of $\alpha = .05$, then, as before, we select the critical value of PRE so that the calculated value of PRE has only a .05 probability of exceeding that critical value if $\eta^2 = 0$. In this case, we see from Figure 4.14 that the probability of obtaining a PRE less than or equal to .20 equals .96, so only 4% of the observed values of PRE should be greater than .20 if $\eta^2 = 0$. Remember that $\eta^2 = 0$ implies that Model C and Model A are identical, so our decision rule will be to reject Model C in favor of Model A if PRE > .20. Now we can use the column for $\eta^2 = .3$ to answer our "what if" question. The entry in that column for the row for PRE = .20 reveals that the probability that the calculated PRE will be less than .20 is .25. That is, even if $\eta^2 = .3$ (i.e., there is a real difference between Model C and Model A), there is still a 25% chance that we will obtain a value of PRE below the critical value and hence will not reject Model C. In other words, the probability of making a Type II error is .25. Conversely, 1 – .25 = .75 is the probability of obtaining PRE > .20 and hence rejecting Model C in favor of MODEL A if $\eta^2 = .3$. In other words, the power (probability of not making a Type II error) is .75.

We can easily do the "what if" analysis for other values of η^2. For example, if $\eta^2 = .05$, which implies a small difference between Models C and A, then power equals 1 – .86 = .14. That is, the chances of obtaining a PRE large enough to reject Model C in favor of Model A would be only 14% for this small difference between the two models. On the other hand, if $\eta^2 = .75$, which implies a very large difference between Models C and A, then power equals 1 – .00 = 1. That is, for this large difference we would be virtually certain to obtain a PRE large enough to reject Model C in favor of Model A.

FIGURE 4.15 Power table for α = .05 when PC = 0 and PA = 1

			Prob(PRE > critical value)							
	Critical values		*True PRE,* η^2							
n	*F*	*PRE*	*0*	*.01*	*.03*	*.05*	*.075*	*.1*	*.2*	*.3*
2	161.45	.994	.05	.05	.05	.05	.05	.06	.06	.07
3	18.51	.903	.05	.05	.05	.06	.06	.07	.08	.11
4	10.13	.771	.05	.05	.06	.06	.07	.08	.11	.15
5	7.71	.658	.05	.05	.06	.07	.08	.09	.14	.21
6	6.61	.569	.05	.05	.06	.07	.09	.10	.17	.26
7	5.99	.499	.05	.06	.07	.08	.10	.12	.20	.31
8	5.59	.444	.05	.06	.07	.09	.11	.13	.23	.36
9	5.32	.399	.05	.06	.08	.09	.12	.14	.26	.41
10	5.12	.362	.05	.06	.08	.10	.13	.16	.29	.46
11	4.96	.332	.05	.06	.08	.11	.14	.17	.32	.50
12	4.84	.306	.05	.06	.09	.11	.15	.18	.35	.54
13	4.75	.283	.05	.06	.09	.12	.16	.20	.38	.58
14	4.67	.264	.05	.06	.09	.13	.17	.21	.41	.62
15	4.60	.247	.05	.07	.10	.13	.18	.23	.44	.66
16	4.54	.232	.05	.07	.10	.14	.19	.24	.46	.69
17	4.49	.219	.05	.07	.10	.14	.20	.25	.49	.72
18	4.45	.208	.05	.07	.11	.15	.21	.27	.52	.74
19	4.41	.197	.05	.07	.11	.16	.22	.28	.54	.77
20	4.38	.187	.05	.07	.12	.16	.23	.29	.56	.79
22	4.32	.171	.05	.07	.12	.18	.25	.32	.61	.83
24	4.28	.157	.05	.08	.13	.19	.27	.35	.65	.87
26	4.24	.145	.05	.08	.14	.20	.29	.37	.69	.89
28	4.21	.135	.05	.08	.15	.22	.31	.40	.72	.92
30	4.18	.126	.05	.08	.15	.23	.33	.42	.75	.93
35	4.13	.108	.05	.09	.17	.26	.37	.48	.82	.96
40	4.09	.095	.05	.10	.19	.29	.42	.54	.87	.98
45	4.06	.085	.05	.10	.21	.32	.46	.59	.91	.99
50	4.04	.076	.05	.11	.23	.36	.51	.64	.93	**
55	4.02	.069	.05	.11	.25	.39	.55	.68	.95	**
60	4.00	.064	.05	.12	.27	.42	.58	.72	.97	**
80	3.96	.048	.05	.14	.34	.53	.71	.84	.99	**
100	3.94	.038	.05	.17	.41	.62	.81	.91	**	**
150	3.90	.026	.05	.23	.57	.80	.93	.98	**	**
200	3.89	.019	.05	.29	.70	.90	.98	**	**	**
500	3.86	.008	.05	.61	.98	**	**	**	**	**

** Power > .995

The cumulative sampling distributions are unwieldy, and only a few of the numbers are actually needed for the "what if" analyses, so a "power table" that only gives the power probabilities for specified levels of η^2 is more useful. Figure 4.15 gives the power probabilities for selected values of η^2 and *n* when PC = 0, PA = 1, and α = .05. We can use this table to do the same "what if" analyses that we did above, as well as many others. For our example problem we simply use the row for *n* = 20: for η^2 = .05, power = .16; and for η^2 = .3, power = .79. (The small differences from the power calculations above are due to using the more precise critical value of .187 for PRE instead of .20.)

The power table allows us to ask another kind of "what if" question: "What would the power be if the sample size were increased?" For example, how much would the power increase if the lottery bids were evaluated with 30 bidders instead of 20? If $\eta^2 = .05$, then power = .23, which is better than .16 for 20 bidders, but still not very good. If $\eta^2 = .3$, then power = .93, which is much higher than .79 for 20 bidders and gives an excellent chance of rejecting Model C in favor of Model A. Note that for $n = 50$ we are virtually certain of rejecting Model C whenever the true PRE η^2 is equal to or greater than .3.

Too many researchers fail to ask "what if" power questions before they collect their data. The consequence is often a study that has virtually no chance of rejecting Model C even if the idea that motivated the study is correct. With power tables such as Figure 4.15 (most software programs, such as R and SAS, have easily accessible procedures for calculating power), asking "what if" power questions is so easy that there is no excuse for not asking those questions before collecting data. A natural question is how high should statistical power be? Cohen (1977) suggested that power should be at least .8. However, the ultimate decision is how much the researcher is willing to accept the risk of not finding a significant result even when the ideas motivating the study are correct.

Now that we know how to answer easily "what if" power questions, we need to know what values of true PRE or η^2 are appropriate for those "what if" questions. There are three ways to obtain an appropriate value for η^2 to use in the power analysis: (a) obtain values of PRE from similar research; (b) use Cohen's (1977) suggested values for "small," "medium," and "large" effects; and (c) compute expectations for PRE from guesses about the parameter values. We consider each in turn.

First, with sufficient experience in a research domain, researchers often know what values of PRE are important or meaningful in that domain. Those values of PRE from experience can be used directly in the power table. For example, if, based on past experience, we thought that important effects (such as the effect of the lottery bids) produced PREs greater than or equal to .1, then we could use the $\eta^2 = .1$ column of the power table. If we wanted to ensure that power > .8, then going down the column we find that the first power > .8 requires a sample size between 60 and 80, probably about 73, which is a much greater number of participants than we included in our test of whether people were willing to pay significantly less than the expected value of a lottery ticket.

In using the results of past studies to select an appropriate η^2 for the "what if" power analysis, we must remember that calculated values of PRE are biased because on average they overestimate η^2. The following simple formula can be used to remove the bias from PRE:

$$\text{Unbiased estimate of } \eta^2 = 1 - (1 - \text{PRE})\left[\frac{n - \text{PC}}{n - \text{PA}}\right]$$

For our example in which we calculated PRE = .209, the unbiased estimate of η^2 equals:

$$1 - (1 - .209)\left[\frac{20 - 0}{20 - 1}\right] = 1 - .791\left[\frac{20}{19}\right] = .167$$

Thus, although the value of PRE calculated from the data is .209, for planning further research our best unbiased guess for the true value of η^2 is only .167. The correction for bias has more of an effect for small values of n – PA than for large values. In essence, the adjustment corrects for the ability of least squares to capitalize on chance for small sample sizes. These unbiased estimates are thus sometimes referred to as "adjusted" values.

A second and related method for finding appropriate values of η^2 is to use the values suggested by Cohen (1977) as "small" ($\eta^2 = .02$), "medium" ($\eta^2 = .13$), and "large" ($\eta^2 = .26$). Our power table does not have columns for these specific values of η^2, but .03, .1, and .3 could be used instead. Although these suggested values for small, medium, and large effects are inherently arbitrary, they do represent experience across a wide range of social science disciplines. If you have sufficient experience in a research domain to consider these suggested values unreasonable, then simply use those values that are reasonable based upon your experience. The goal of a power analysis conducted before the collection of data is not an exact calculation of the statistical power but an indication of whether there is much hope for detecting the effect you want to find with the sample size you have planned. If a study would not have much chance of distinguishing between Model C and Model A for a large effect ($\eta^2 = .26$ or .3), then there is little if any reason for conducting the study.

As an example of this approach, let us estimate the power for detecting small, medium, and large effects for the lottery bids using our sample of 20 bidders. Using the row of the power table for $n = 20$ and the columns for $\eta^2 = .03$, .1, and .3, we find that the respective powers are .12, .29, and .79. In other words, we would not have much chance of detecting small and medium effects but a decent chance of detecting a large effect. If we wanted to be able to detect medium effects, then we would need to increase the number of participants in our study.

A third approach to finding an appropriate value of η^2 to use in "what if" power analyses involves guesses about the parameter values and variance. To have reasonable expectations about the parameter values and variance generally requires as much or more experience in a research domain as is necessary to know typical values of PRE. Thus, this third approach is generally less useful than the first two. We present this approach in order to be complete and because the derivation of this approach provides further useful insights about the meaning of PRE and η^2. Also, this approach requires describing in detail the data that one expects to obtain, and such an exercise can often be useful for identifying flawed research designs.

We begin with our familiar definition of PRE:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{\text{SSR}}{\text{SSE(C)}}$$

We have noted before that SSE(C) = SSE(A) + SSR (i.e., the error for the compact model includes all the error of the augmented model plus the error that was reduced by the addition of the extra parameters in the augmented model). Hence, substituting for SSE(C) yields:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(A)} + \text{SSR}} = \frac{1}{\text{SSE(A)/SSR} + 1}$$

To obtain a definition of the true proportional reduction in error η^2, we simply estimate SSE(A) and SSR, using not the data but the parameter values of B_0 and β_0 that are of interest.

For example, we noted above that:

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{Y}_{iC} - \hat{Y}_{iA})^2$$

If we thought that the effect of the lottery bids would be to decrease the mean bid from 50 to 40, then $\hat{Y}_{iC} = B_0 = 50$ represents the null hypothesis and $\hat{Y}_{iA} = \beta_0 = 40$ represents an alternative hypothesis that we want to test. For that situation we would *expect*:

$$\mathrm{SSR} = \sum_{i=1}^{20} (50 - 40)^2 = \sum_{i=1}^{20} 10^2 = \sum_{i=1}^{20} 100 = 2000$$

In other words, the SSR that we expect is simply 100 added up 20 times (once for each bidder). We saw in Chapter 2 that SSE/(n – PA) was an estimate of the variance σ^2. If we use our expected value of $\beta_0 = 40$ to calculate SSE(A), then we are not using data to estimate any parameters, so PA = 0. Hence, SSE(A)/n = σ^2, so the value of SSE(A) that we *expect* to obtain is:

$$\mathrm{SSE(A)} = n\sigma^2$$

Thus, if we have a reasonable guess or expectation for the variance, then we can easily calculate the value of SSE(A) that we would expect. Having good intuitions about what variance to expect is usually as difficult or more difficult than knowing what PRE to expect. Both are based on previous experience in a research domain. Good guesses for the variance often depend on previous experience with the particular measure for Y. For our example, suppose that past data for the lottery lead us to expect that σ^2 is about 400; then, we would expect:

$$\mathrm{SSE(A)} = n\sigma^2 = (20)400 = 8000$$

We now can return to our formula for PRE to calculate the value that we expect for the true proportional reduction in error η^2 (given our guesses for B_0, β_0, and σ^2).

$$\text{Expected } \eta^2 = \frac{1}{\mathrm{SSE(A)/SSR} + 1} = \frac{1}{8000/2000 + 1} = \frac{1}{4 + 1} = .2$$

In other words, $\eta^2 = .2$ corresponds to our guesses about B_0, β_0, and σ^2. We now can use the power tables to find the power we would have for comparing Model C, which predicts that the mean bid will be 50, against Model A, which predicts that the mean bid will be 40, when the variance is about 400. Using Figure 4.15, we find that for 20 observations the probability of rejecting Model C (i.e., deciding that the lottery bids were less than the expected value of a ticket) is only about .56 even if we think that it will on average be $10 less than the expected value of $50. This means that the researcher has only a little more than a 50/50 chance of *deciding* that lottery bids are lower than the expected value of a ticket even when they really are lower by $10. Using the power table, we can see that testing our hypothesis with twice the number of persons would increase the power substantially to .87. In this case, it would seem advisable to test our hypothesis with a larger sample size rather than with a sample size that offered little hope of finding the effect. We can similarly calculate the power for other "what if" values of B_0, β_0, and σ^2 that we might want to consider.

IMPROVING POWER

The relatively low power—the high probability of making a Type II error—for the apparently reasonable evaluation of the hypothesis we have been considering may be startling. Unfortunately, low power is a problem that plagues data analysis far more frequently than is commonly realized. Low power creates serious difficulties. For example, consider the plight of researchers trying to evaluate the effectiveness of an innovative educational curriculum or a new therapy for a serious illness. If the power of detecting the effect is only about 50/50, there is a fairly high risk of concluding that the new curriculum or therapy is not effective even when it actually is. When one considers the time and money that may be invested in research—not to mention the potential of the findings to benefit people and advance science—it generally makes little sense to design and conduct studies that have little chance of finding effects even when the effects really exist. In our particular example, the obtained value of PRE allowed us to reject the hypothesis that $\beta_0 = 50$; we were either lucky, or the true value of β_0 was considerably less than the alternative value of 50 that we considered above. In general, however, we want to increase the power of the statistical inference. There are three basic strategies for improving power: reduce error, increase α, and/or increase the number of observations. We consider each in turn.

Reducing Error

One way to reduce error is to control as many of the possible random perturbations as possible. In our lottery example, one might reduce error and obtain more power by providing clear instructions to participants, making sure participants were well rested, eliminating distractions in the bidding environment, and using a more reliable bidding procedure. In other words, error is reduced by obtaining data of better quality. In the equation:

DATA = MODEL + ERROR

the model will account for a higher proportion of the data if the data are of higher quality and hence have less error. Less error allows us to obtain a more powerful look at our data. Although reducing error by such means may be the most effective method for improving power, the techniques for doing so are usually domain-specific and outside the scope of this book.

Another way to reduce error is to improve the quality of the model. Again in the equation:

DATA = MODEL + ERROR

error will be smaller for data of fixed quality if the model can be improved to account for more of the data. How to use models more complex than the simple models we have been considering in these beginning chapters is the subject of the remainder of the book, so we cannot give too many details here. The general idea is to build models that make predictions conditional on additional information we have about the observations. In the lottery example we might know, say, which bidders had participated in lotteries before and which ones, if any, had ever won a lottery. If having participated in a lottery before makes a difference in the amount an individual is willing to pay for a lottery ticket, then we can make different predictions conditional on whether (or perhaps how often) the bidder had participated in lotteries previously. By doing so we will have, in essence, removed what was formerly a random perturbation—previous participation in lotteries—from the error and included it in the model. Again the reduced error will give us a more powerful look at our data. In later chapters we explicitly consider the addition of parameters to the model for the purpose of improving power.

Increasing α

A different way to improve power is to increase α, the probability of a Type I error. The probabilities of Type I and II errors are linked in that if we choose a critical value that increases (decreases) α then we simultaneously and unavoidably decrease (increase) the probability of a Type II error. For our lottery data with $n = 20$, Figure 4.16 shows power as a function of η^2 and α. As α increases from .001 to .25 the critical values for F and PRE decrease. It obviously becomes easier for the values of F and PRE calculated from the data to beat these critical values, so the power increases as α increases. For example, if we do a "what if" power analysis with $\eta^2 = .2$, then the power at $\alpha = .05$ is .56, but if we increase α to .1, then power increases to .70.

Editors of scientific journals are wary of Type I errors and will seldom accept the use of $\alpha > .05$ for statistical inference. However, there are many practical data analysis problems when a higher α is justified to increase power. Characteristics of such data analyses are (a) that increasing power in any other way is infeasible, (b) that rejection of Model C if it is indeed false would have important practical consequences, and (c) that the costs associated with a Type I error are not great. Consider, for example, the statistical decision problem faced by a researcher testing the effectiveness of an innovative new curriculum. It may be difficult for the researcher to control any other sources of error to increase power. It would certainly be important to identify a curriculum that could improve student learning. The consequence of a Type I error would probably be further trial use of the curriculum in several classrooms the following year, which may not involve a significant cost. Thus, the researcher might well adopt a higher α to choose the critical values for PRE and F for her statistical inference.

Conversely, note that reducing α also reduces power. In the lottery example, lowering α to .01 would increase our protection against the possibility of a Type I error but would

FIGURE 4.16 Power for PA = 1, PC = 0, and n = 20 for various levels of α

			Prob(PRE > critical value)							
	Critical values		*True PRE, η^2*							
α	*F*	*PRE*	*0*	*.01*	*.03*	*.05*	*.075*	*.1*	*.2*	*.3*
.001	15.08	.443	.00	.00	.00	.01	.01	.02	.09	.22
.005	10.07	.346	.01	.01	.02	.03	.05	.07	.21	.43
.01	8.18	.301	.01	.02	.03	.05	.08	.11	.30	.54
.025	5.92	.238	.03	.04	.07	.10	.15	.20	.44	.69
.05	4.38	.187	.05	.07	.12	.16	.23	.29	.56	.79
.1	2.99	.136	.10	.13	.20	.26	.34	.42	.70	.88
.25	1.41	.069	.25	.29	.38	.46	.55	.63	.85	.96

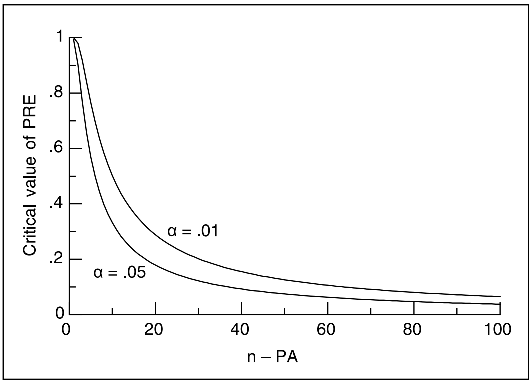
reduce power to less than .30, a level for which it would not likely be worth conducting the study. If the costs of a Type I error are very high—for example, if reliance on the research findings involved extensive teacher retraining in the case of an innovative new curriculum—then the use of restrictive values of α may be appropriate. However, in some cases, such as when the sample size is very small, as it is for our lottery study, reducing α would reduce the power so much that it would no longer be worthwhile to do the study—the chances of seeing anything are too low with such a low-powered "microscope."

Increasing *n*

Probably the most common technique for increasing the power of statistical tests is to increase the number of observations. Figure 4.17 is a graphical display of the 95% (α = .05) and 99% (α = .01) critical values of Figure 4.9 as a function of the number of observations. The value of PRE required to reject Model C drops dramatically as the number of observations increases. For example, had there been 51 participants instead of 20 in our lottery study, a PRE of only .075 would have been required (α = .05) to reject the Model C that assumed B_0 = 50 instead of the PRE of .187 required in our example. The drop in the critical values for F and PRE needed to reject Model C corresponds to an increase in power, as we have noted several times. For example, for η^2 = .2 as a "what if" value of the true proportional reduction in error, the power for 20 observations is .54 (see Figure 4.15), but with 51 observations the power increases to .93.

There are two reasons for not routinely using a large number of observations. First, it may be infeasible, due to cost or other data collection constraints, to obtain more observations. Second, power might be so high that some statistically significant results may be misleading. By turning up the power on our metaphorical statistical microscope to extraordinary levels, we might detect flaws in Model C that are statistically reliable but are trivial substantively. For example, with 120 observations any PRE greater than .032 is cause to reject Model C in favor of Model A. However, the 3.2% reduction in error means that Model A may be a trivial improvement over Model C. For this reason, one should always report not just the statistical inference about whether Model C is or is not rejected but also the obtained values for PRE and F so that the reader can evaluate

FIGURE 4.17 Critical values of PRE as a function of the number of observations



the magnitude by which Model A improves on Model C. All else being equal, more statistical power is always better; we just need to be careful in interpreting the results as substantively important when power is very high.

It is also important not to dismiss a small but reliable PRE just because it is small. Rejecting Model C in favor of Model A may be important theoretically even if Model A provides only a slight improvement. Whether a given PRE is substantively interesting will depend on theory and prior research experience in the particular domain.

CONFIDENCE INTERVALS

Confidence intervals provide an alternative way for considering statistical inference. Although, as we shall see later, confidence intervals are exactly equivalent to statistical inference as described above, they reorganize the information in a way that can give useful insights about the data and our model.

A *confidence interval* simply consists of all those possible values of a parameter that, when used as a hypothesis for Model C, would not cause us to reject Model C. For example, the mean for the 20 lottery bids is 42.5 and estimates the parameter β_0 in Model A. We have already determined that 50 is not in the confidence interval for β_0 because, when we used 50 as the prediction for Model C, we obtained an unlikely PRE value that would occur less than 5% of the time. Any value greater than 50 would produce an even larger PRE, so none of those values is included in the confidence interval. Conceptually, we could find the boundary of the confidence interval by trying increasingly higher values (i.e., higher than b_0) for B_0 in Model C until we found a value for B_0 that produced a PRE that would not cause us to reject Model C. For example, if we tried $B_0 = 45$, the respective sums of squared errors for Models C and A would be 4392 and 4267, yielding PRE = .028, which is below the critical value for $\alpha = .05$; hence 45 is in the confidence interval for β_0, and the boundary must be somewhere between 45 and 50. We would also need to search for the lower boundary below the estimated value of 42.5. We can avoid this iterative search because it can be shown that the boundaries are given by:

$$b_0 \pm \sqrt{\frac{F_{\text{crit};1,\, n-1;\alpha}\text{MSE}}{n}} \tag{4.4}$$

For the simple model of the lottery bids, MSE = s^2 = SSE/(n – 1) = 4267/19 = 224.58. $F_{\text{crit};1;n-1;\alpha}$ is the critical value at level α for F with 1 degree of freedom for the numerator (i.e., the difference in the number of parameters between the two models) and $n - 1$ degrees of freedom for the denominator (i.e., the number of observations minus the number of parameters in Model A). For $\alpha = .05$, the critical value of $F_{1,19;.05} = 4.38$. For these data, b_0 = the mean = 42.5, so the boundaries of the confidence interval are given by:

$$42.5 \pm \sqrt{\frac{(4.38)224.58}{20}} \quad \text{or} \quad 42.5 \pm 7.01$$

Thus, the lower boundary is 42.5 – 7.01 = 35.49 and the upper boundary is 42.5 + 7.01 = 49.51. We are therefore 95% (1 – α) confident that the true value for β_0 is in the interval [35.49, 49.51]. Any B_0 not in this interval that we might try for Model C would, with the present data, produce values of PRE and F that would lead us to reject Model C. Any B_0 in this interval would produce values of PRE and F below the critical values, so we would not reject Model C.

When estimating statistical power we described how to do an a priori power analysis. Those "what if" power analyses are necessarily inexact. Confidence intervals are useful for describing post hoc the actual statistical power achieved in terms of the precision of the parameter estimates. Wide confidence intervals represent low statistical power and narrow intervals represent high statistical power.

EQUIVALENCE TO THE *t*-TEST

In this optional section we demonstrate the equivalence between the statistical test for the simple model developed in this chapter and the traditional *one-sample t-test* presented in most statistics textbooks. We do so to allow readers with exposure to traditional textbooks to make the comparison between approaches.

The one-sample *t*-test answers whether the mean of a set of observations equals a particular value specified by the null hypothesis. The formula for the one-sample *t*-test is:

$$t_{n-1} = \frac{\sqrt{n}(\bar{Y} - B_0)}{s}$$

where n is the number of observations, $\bar{Y}$ is the calculated mean, B_0 is the value specified by the null hypothesis, and s is the standard deviation of the set of observations. With the appropriate assumptions about Y—the same assumptions that we made about the distribution and independence of the ε_i—calculated values of t can be compared to critical values of *Student's t-distribution*. Tables of this distribution are available in many statistics textbooks. However, separate tables are not really needed because squaring t with $n - 1$ degrees of freedom yields an F with 1 and $n - 1$ degrees of freedom. Thus, the F tables in the Appendix may readily be used.

To show the equivalence between F as presented in this chapter and the usual t-test, we begin with the definition of F for the simple model; that is:

$$F_{1,n-1} = \frac{\text{PRE}/1}{(1 - \text{PRE})/(n - 1)}$$

We know that PRE = SSR/SSE(C), and it is easy to show that:

$$1 - \text{PRE} = 1 - \frac{\text{SSR}}{\text{SSE(C)}} = \frac{\text{SSE(C)} - \text{SSR}}{\text{SSE(C)}} = \frac{\text{SSE(A)}}{\text{SSE(C)}}$$

Substituting these values into the definition for F yields:

$$F_{1,n-1} = \frac{\text{SSR/SSE(C)}}{[\text{SSE(A)/SSE(C)}]/(n - 1)} = \frac{\text{SSR}}{\text{SSE(A)}/(n - 1)}$$

But from Equation 4.2 we know that for the simple model SSR can be replaced with $n(B_0 - \bar{Y})^2$, and from Chapter 2 we know that SSE(A)/(n – 1) is s^2, the variance of the set of observations. Substituting these values yields:

$$F_{1,n-1} = \frac{n(B_0 - \bar{Y})^2}{s^2}$$

Taking the square root of this last equation gives the final result of:

$$\sqrt{F_{1,n-1}} = t_{n-1} = \frac{\sqrt{n}(\bar{Y} - B_0)}{s}$$

We provide the above derivation not to present yet another computational formula but to show that our model comparison approach to statistical inference is statistically identical to the traditional approach. The use of PRE and F for comparing models is nothing more than a repackaging of the traditional approach. This repackaging has the important consequence of making it easy to generalize to more complicated models and data analysis questions. We will use PRE and F just as we did in this chapter for statistical inference throughout the remainder of the book. In contrast, the traditional t-test does not generalize nearly so easily. Also, even though the t-test must produce exactly the same conclusion with respect to the null hypothesis, it does not automatically provide a measure of the magnitude of the result. In our model comparison approach, PRE automatically provides a useful measure of the magnitude.

AN EXAMPLE

In this section we illustrate the techniques of this chapter using the internet access data that were presented in Figure 1.1. Suppose that marketing researchers had projected that 75% of households would have internet access by the year 2013. Was the marketing researchers' projection overly optimistic? The question is equivalent to comparing the following two models:

MODEL A: $Y_i = \beta_0 + \varepsilon_i$

MODEL C: $Y_i = 75 + \varepsilon_i$

We know that the mean is $\bar{Y}$ = 72.806 (see Chapter 2), so the estimated Model A is:

$\hat{Y}_i = 72.806$

For the statistical inference we need to calculate PRE and F from SSE(A) and SSE(C). We also know that the variance or the MSE is 27.654 (see Chapter 2). Since MSE = SSE(A)/(n – 1), we can easily obtain SSE(A) by multiplying MSE by n – 1; thus, SSE(A) = 27.654 (49) = 1355.046; this value (within rounding error) is also given in Figure 2.10. We can compute SSR using:

$$\text{SSR} = n(B_0 - \bar{Y})^2 = 50(75 - 72.806)^2 = 240.682$$

Then it is easy to get SSE(C) from:

$$\text{SSE(C)} = \text{SSE(A)} + \text{SSR} = 1355.046 + 240.682 = 1595.728$$

The computations of PRE and F are then easy:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(C)}} = \frac{240.682}{1595.728} = .151$$

and

$$F_{1,49} = \frac{\text{PRE}/1}{(1 - \text{PRE})/(n - 1)} = \frac{.151}{.849/49} = 8.71$$

From the tables in the Appendix, the critical values for PRE and F (for $\alpha = .05$) are, respectively, about .075 and 4.03. The obtained values clearly exceed the critical values, so we can reject Model C in favor of Model A. Thus, the 15.1% reduction in error obtained by using the estimate $b_0 = 72.806$ instead of the null hypothesis value of $B_0 = 75$ is statistically significant. We can therefore conclude that the percentage of households that had internet access was significantly lower than the marketing researchers' projection. We might summarize our results for a journal article as follows:

> On average, across states the percentage of households that had internet access in the year 2013 ($M = 72.806$) was significantly lower than the projected value of 75%, PRE = .151, $F(1, 49) = 8.71$, $p < .05$.

From the above it is also easy to calculate the 95% confidence interval for β_0, the true average percentage of households that had internet access across states. Substituting the appropriate values into Equation 4.4 yields:

$$72.806 \pm \sqrt{\frac{4.03(27.654)}{50}} \quad \text{or} \quad 72.806 \pm 1.493$$

which gives an interval of [71.313, 74.299]. Using this interval, we can easily ask other questions. For example, had the marketing researchers projected that $B_0 = 73$, we would not conclude that the actual percentage of households with internet access was significantly less than the projection, because 73 is included in the 95% confidence interval.

SUMMARY

In Chapter 1 we noted that the equation:

DATA = MODEL + ERROR

implies an inherent tension in data analysis between reducing error as much as possible and keeping the model as simple or parsimonious as possible. Whenever we consider adding an additional parameter to the model so that it will fit the data better and thereby reduce error, we must ask whether the additional complexity of the model is worth it. In this chapter we have developed inferential machinery for answering whether the additional complexity is worth it.

To decide whether the benefits of the additional parameters in Model A outweigh the benefits of the parsimony and simplicity of Model C, we first calculate SSE(A) and

SSE(C), respectively, the sum of squared errors for the augmented model (which incorporates the additional parameters) and the compact model (which does not include those parameters). The sum of squares reduced, SSR, is simply the difference between them:

$$\text{SSR} = \text{SSE(C)} - \text{SSE(A)}$$

Then we calculate the proportional reduction in error attributable to the additional parameters, which is given by:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{\text{SSR}}{\text{SSE(C)}}$$

Another related statistic is the ratio of the proportional reduction in error per parameter added to the potential proportional reduction in error per remaining unused parameter, which is given by:

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

We then compare the calculated values of PRE and *F* to the distribution of values we would expect *if* Model C, the compact model, *were true*. If the calculated values of PRE and *F* would have been unlikely if Model C were true, then we reject Model C and conclude that the extra complexity of Model A is worth it. On the other hand, if the calculated values are ones that might reasonably have been obtained if Model C were true, then we do not reject Model C and without further evidence we would not accept the additional complexity of Model A.

This inferential machinery is merely a guide for decision making and is not infallible. There are two kinds of errors that we can make. A Type I error occurs when Model C is in fact correct but by chance we happen to get unusual values of PRE and *F* and so reject Model C. The probability of a Type I error is α and defines how unusual PRE and *F* have to be before we reject Model C. A Type II error occurs when Model C is in fact false or inferior to Model A but by chance we happen to get values of PRE and *F* that are not unusual and so fail to reject Model C. We generally select α, the probability of a Type I error, and try to reduce the chances of a Type II error by collecting better data with less error and by increasing the number of observations. Reducing the chances of a Type II error is referred to as increasing the statistical power of an inference.

We developed this inferential machinery in the context of asking a question for the simple model. However, *exactly* the same procedure will work for all the more complex models we consider in subsequent chapters. In this chapter, we have learned all we need to know as data analysts about statistical inference. The remainder of our task is to learn how to build more complex and interesting models of our data.