CHAPTER 2

Predictions and Projections: Some Issues of Research Design

"There will be no nuclear war within the next fifty years."

"In the period 1965–70, Mao Tse-tung and De Gaulle will die."

"Major fighting in Viet-Nam will peter out about 1967; and most objective observers will regard it as a substantial American victory."

"In the United States Lyndon Johnson will have been re-elected in 1968."

—Ithiel de Sola Pool[1]

Introduction

Projections of the future can be useful or embarrassing, depending on their accuracy. The assumption that a wide range of factors remain constant or continue to change at current rates can quickly crumble.[2] And yet how imbedded in our thought is the idea that the future is a straightforward projection of the past: we may doubt the optimism of Professor Pool's first prediction if only because of the failure of the other predictions on the list. At least, unlike some predictions, these have the modest virtue of being explicit, and it is easy to tell whether they went wrong.[3]

[1] "The International System in the Next Half Century," in Daniel Bell, ed., *Toward the Year 2000: Work in Progress* (Boston: Beacon Press, 1967), pp. 319–20.

[2] A very useful discussion of the assumptions behind many projections is Otis Dudley Duncan, "Social Forecasting—The State of the Art," *The Public Interest,* no. 17 (Fall 1969), 88–118.

[3] On previous prophecies, see Arthur M. Schlesinger, "Casting the National Horoscope," *Proceedings of the American Antiquarian Society,* 55 (1945), 53–93.

Almost all efforts at data analysis seek, at some point, to generalize the results and extend the reach of the conclusions beyond a particular set of data. The inferential leap may be from past experiences to future ones, from a sample of a population to the whole population, or from a narrow range of a variable to a wider range. The real difficulty is in deciding when the extrapolation beyond the range

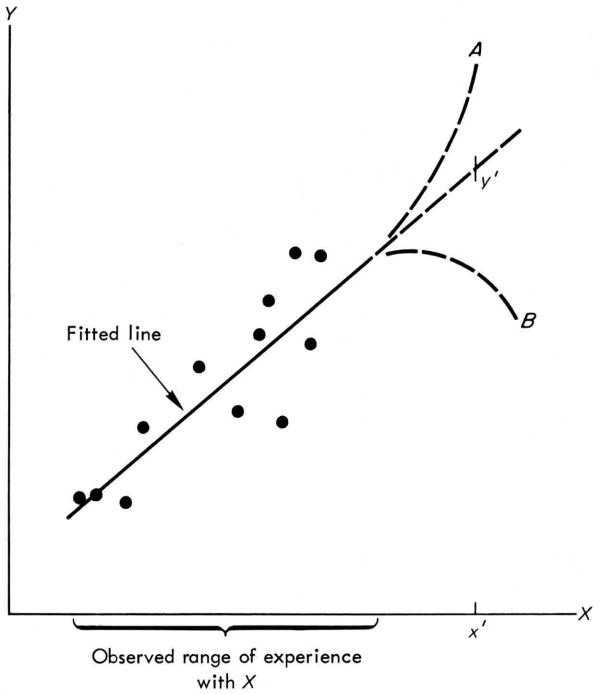


FIGURE 2-1 Problem of simple extrapolation

Q: Should the fitted line be extended to predict the value y' for the new observation x' (which is outside the range of previous experience with the x-variable)? Or, is A or B a better model?

A: "*A priori* nonstatistical considerations . . ."

of the variables is warranted and when it is merely naive. As usual, it is largely a matter of substantive judgment—or, as it is sometimes more delicately put, a matter of "a priori nonstatistical considerations" (Figure 2-1).

If the observed variation in a variable is small relative to its total possible variation, then the extension of the inference based on a narrow range of observations is less warranted than extrapolation

based on a wider range of observed variations. Equally obvious is the observation that the risk of error is less if the extrapolated value is "close" to the previous pattern of experience rather than greatly different, other things being equal. In some cases it may be useful to conduct trial runs at extrapolation by using a fraction of the available data to produce a fitted curve, using the remaining data to test the accuracy of the extrapolated results. Obviously if the conditions governing a relationship change in relevant respects, the effort at extension of results is in danger of making errors.

Simple extrapolation involves the extension of results outside the range of experience of a single describing variable. A more subtle situation arises in the multivariate case involving extrapolation beyond the range of the *combination* of experience jointly observed in two or more describing variables. Karl A. Fox has described this situation as "hidden extrapolation."[4]

Figure 2-2 shows the pattern of correlation between two describing variables. Assume these two describing variables, X_1 and X_2, are used in combination to predict a response variable, Y. The situation appears to be relatively satisfactory because there is a wide range of experience with both X_1 and X_2. But note how little experience there is concerning certain *combinations* of X_1 and X_2—since all the points representing joint occurrences of X_1 and X_2 are contained in the narrow band surrounding the line. There is no experience with combinations such as low X_1-high X_2 (in the upper left of the rectangle) or high X_1-low X_2 (lower right) and how such unobserved combinations of X_1 and X_2 might affect the response variable. The response variable may behave very differently for such combinations of X_1 and X_2. Thus a prediction equation, predicting Y from X_1 and X_2, may be quite misleading if applied to situations in which X_1 and X_2 occur in combinations different from those observed here.

Thus the extension of the inference over all combinations of X_1 and X_2 may founder on the possibility of an interaction effect between X_1 and X_2 in their influence on Y in the region of the combinations with which there is no experience. The problem arises because of limited experience with the *joint* relationship of X_1 and X_2, even though there may be extensive experience with the entire range of each variable taken singly. Thus the name, "hidden extrapolation."

The problem arises in any predictive study involving correlated describing variables. Figure 2-3 shows the narrowed range of joint experience in the case of three correlated describing variables.

We diagnose the problem by considering the scatterplots of the

[4]This discussion is based on Karl A. Fox, *Intermediate Economic Statistics* (New York: Wiley, 1968), pp. 265–66.

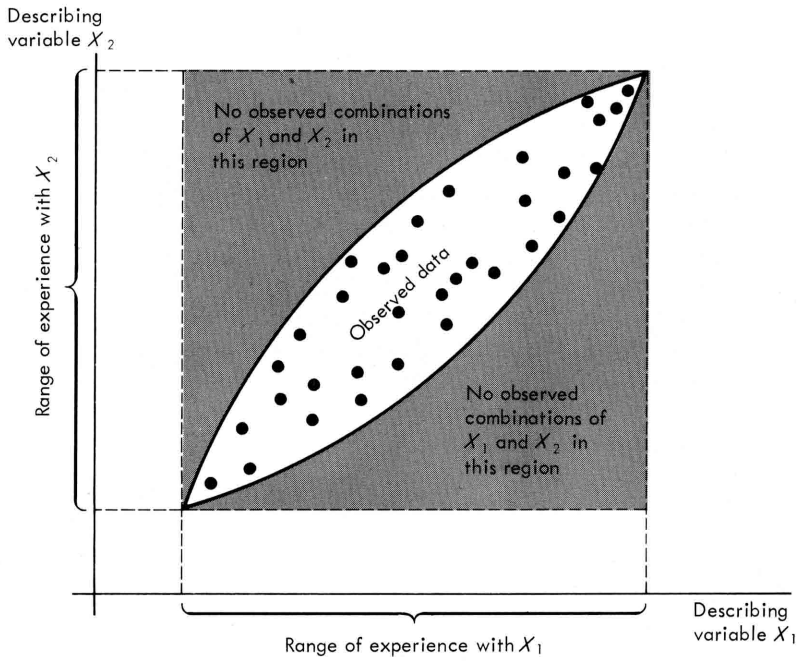


FIGURE 2-2 Correlation between two describing variables

relationships between the describing variables and by looking over the original joint observations. Cures for the difficulty include the collection of additional data, particularly of "deviant cases" in areas outside the previously experienced combinations of describing variables.

Let us now turn to several examples illustrating and evaluating methods of prediction. These case studies show different statistical tools in action. Note, however, that the central consideration in most cases is the research design, rather than the mechanics of using the statistical tool. Mosteller and Bush make this point quite sharply:

> We first wish to emphasize that formal statistics provides the investigator with tools useful in conducting thoughtful research; these tools are not a substitute for either thinking or working. A major goal for the statistical training of students should be statistical thinking rather than statistical formulas, by which we mean specifically: thinking about (1) the conception and design of the study and what it is that is to be measured and why, (2) the definitions of the terms being used, and how modifications in definition might change both the outcome and the interpretation of a study, (3) sources of variation in every part of the study, including such things as

individual differences, group and race differences, environmental differences, instrumental or measuring errors, and intrinsic variation fundamental to the process under investigation. In no circumstances do we think that sophisticated analytical devices should replace clean design and careful execution, unless very unusual economic considerations arise. However, it may be worth remarking that crude data collected as best the investigator could may require the most advanced statistical tools. Here a quotation from Wallis may be appropriate:

> In general, if a statistical investigation . . . is well planned and the data properly collected the interpretation will pretty well take care of itself. So-called "high-powered," "refined," or "elaborate" statistical techniques are generally called for when the data are crude and inadequate—exactly the opposite, if I may be permitted an obiter dictum, of what crude and inadequate statisticians usually think."[5]

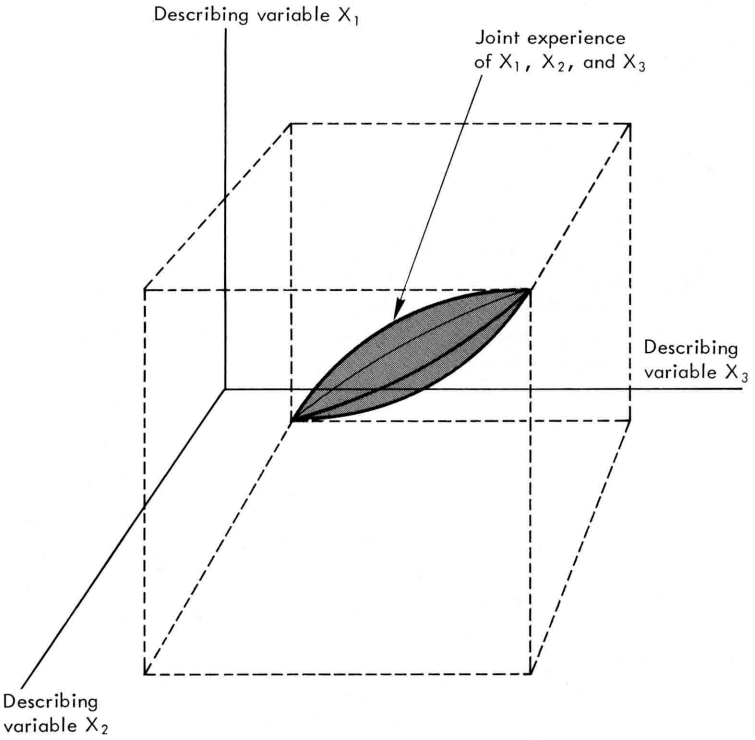


FIGURE 2-3 Range of joint experience—three describing variables

[5] Frederick Mosteller and Robert R. Bush, "Selected Quantitative Techniques," in Gardner Lindzey, ed., *Handbook of Social Psychology:* Vol. I, *Theory and Method* (Cambridge, Mass.: Addison-Wesley, 1954), p. 331. The passage by Wallis is found in W. Allen Wallis, "Statistics of the Kinsey Report," *Journal of the American Statistical Association,* 44 (1949), p. 471.

Problem in Prediction: The National Crime Test and a Cancer Test

Assessing the quality of a prediction or extrapolation can sometimes be a tricky matter. Consider the following example, which reveals the interplay between the properties of the predictive device and the tested population.

A proposal was once made that every 6-, 7-, and 8-year-old child (a total of 13 million in all) be given psychological tests to identify potential "criminality" in order that the supposed lawbreakers of the future be given some sort of treatment. The proposal encountered a storm of moral, legal, and technical criticism which led to its apparent abandonment. One of the technical flaws, which also serves to emphasize the moral and legal criticism of the proposal, is shown in the following model. Assume the National Crime Test has the following hypothetical properties:

1. It will successfully identify 40 percent of those arrested in the future. (Unfortunately, a child's "identification" by the NCT might help insure his future arrest through the mechanism of a self-fulfilling prophecy, operating with respect to the child or the police or both. Perhaps even NCT scores would be used to convince a jury of the guilt of the accused—thereby further increasing the "accuracy" of the prediction.)
2. It will also correctly classify 90 percent of those children who will not be arrested in the future.

Do these characteristics of our hypothetical NCT indicate it is a useful predictor of criminality? It might seem so, since it does identify four out of ten of the future "bad guys" and nine out of ten of the "good guys." But let us look into the errors in prediction made by a test with these characteristics. Assuming that three percent of these children will, later in life, commit a serious crime, we can construct Table 2-1, which shows the predictive performance of the NCT.

The table shows the errors made in the test; let us consider the "false positives" in which the test predicts criminality incorrectly. The upper righthand corner of the table shows 1,261,000 false positives compared to 156,000 correct predictions of criminality. Thus for every correct prediction of future difficulties, there are eight incorrect ones! In this light, such a test would be unacceptable to most people—even though its predictive characteristics, as originally expressed, seemed impressive. Furthermore, the assumptions we made about the predictive powers of such tests were, if anything, much too generous, given the poor performance of psychological tests of "criminality."

TABLE 2-1
Hypothetical (Fortunately) National Crime Test

		Reality	
		Criminal	*Noncriminal*
Test predicts	*Criminal*	156,000	1,261,000
	Noncriminal	234,000	11,349,000
		390,000	12,610,000
		Total = 13,000,000	

COMPUTATIONS:
3 percent of 13,000,000 children will commit a serious crime:
(.03)(13,000,000) = 390,000 children. NCT accurately predicts 40 percent:
(.40)(390,000) = 156,000
97 percent of 13,000,000 are not future criminals:
(.97)(13,000,000) = 12,610,000. NCT accurately predicts 90 percent:
(.90)(12,610,000) = 11,349,000.

Consider another example of the same problem. A hypothetical test for cancer has the following characteristics:

1. Pr(test positive | cancer) = .95. This conditional probability indicates that the test reads "positive" 95 percent of the time given that the person tested in fact has cancer.
2. Pr(test negative | no cancer) = .96.

In other words, the test correctly identifies, on the average, 95 out of 100 of those who do have cancer and also 96 out of 100 of those who do not have cancer. These characteristics give the following table of probabilities:

		Reality	
		Cancer	*No cancer*
Test predicts	*Positive*	.95	.04
	Negative	.05	.96
		1.00	1.00

Now assume that one percent of those tested actually do have cancer; that is, Pr(cancer) = .01. (This is an unconditional probability, since it depends upon no given prior condition.) Note that since only one percent of those tested have cancer, the flow of those tested is mainly down the righthand column of the table of probabilities.

What proportion of false positives (and false negatives) will be

TABLE 2-2
Computation of Probabilities

We have the following data:
Pr(cancer) = .01
Therefore Pr(not cancer) = 1.00 − .01 = .99.
Similarly,
Pr(test positive | cancer) = .95, and therefore
Pr(test negative | cancer) = .05.
Also,
Pr(test negative | no cancer) = .96, and therefore
Pr(test positive | no cancer) = .04.

The problem is to compute Pr(cancer | test positive), which equals, by Bayes' theorem:

$$\frac{\Pr(\text{test positive} \mid \text{cancer})\,\Pr(\text{cancer})}{\Pr(\text{test positive} \mid \text{cancer})\,\Pr(\text{cancer}) + \Pr(\text{test positive} \mid \text{not cancer})\,\Pr(\text{not cancer})}$$
$$= \frac{(.95)(.01)}{(.95)(.01) + (.04)(.96)} = .19.$$

produced by the test? One way to answer with respect to false positives is to compute Pr(cancer | test positive)—the probability that a person has cancer, given that the test reads positive. This can be done, using the appropriate equations for conditional probabilities, shown in Table 2-2. Another way to handle the problem is to consider what happens when, say, 10,000 people are screened for cancer using the hypothetical test. Computations analogous to those in Table 2-1 yield the following expected results:

		Reality	
		Cancer	*No cancer*
Test predicts	*Positive*	95	396
	Negative	5	9,504

and therefore

$$\Pr(\text{cancer} \mid \text{positive}) = \frac{95}{95 + 396} = .19.$$

Thus about 19 percent of those indicated positive will actually have cancer; 81 percent of the positives will be false. The decision whether this is a good test depends upon the cost of such false positives and their consequent detection as well as the benefits that derive from

the detection of the disease. Perhaps such a test would be most useful as a screening device to indicate patients needing further tests.

Similar arguments apply to the use of lie detectors, the prediction of juvenile delinquency on the basis of family background, and the use of "preventive detention."[6] The reason the original qualities of the prediction seem to collapse when the test is applied to data is that, in these two cases, the quality to be detected is rather rare. Therefore, even though the hypothetical cancer test correctly predicts cancer 95 percent of the time and noncancer 96 percent of the time, so many people (99 percent in our example) flow through the right (noncancer) side of the table of probabilities that even the low error *rate* (4 percent) produces a large *number* of errors *relative* to the number of correct predictions of cancer. If, on the other hand, *half* the tested population had cancer, then the expected table (for 10,000 people) would be:

		Reality	
		Cancer	*No Cancer*
Test predicts	*Positive*	4750	200
	Negative	250	4800

This is pretty sensational predicting!

The *properties of the test are the same* in both cases, but the populations tested differ with respect to the distribution of the characteristic to be detected. Thus a test which does a good job of prediction on one population may not perform so well on a second trial if distribution of the characteristic sought differs markedly in the second population. Thus it will be worthwhile to try out—if only by working through the arithmetic as we have done here—the test on a population for which the distribution of the characteristic to be predicted is the same as the population for which the ultimate prediction is to be made. Note that the two numbers Pr(positive | cancer) and Pr(negative | not cancer) were not enough to describe adequately the performance of the prediction. Instead, a third piece of information, in this case Pr(cancer), was necessary to permit an adequate assessment of the performance of the test for that population.

[6]See Jerome H. Skolnick, "Scientific Theory and Scientific Evidence: An Analysis of Lie-Detection," *Yale Law Journal*, 70 (April 1961), 694–728; and Travis Hirschi and Hanan C. Selvin, *Delinquency Research* (New York: Free Press, 1967), chap. 14.

Finally, some very high rates of successful "prediction" should not fool us. After all, we can achieve 99 percent "accuracy" simply by predicting that no person has cancer. Since 99 percent of the people in our example don't have cancer, the rule is 99 percent "accurate" in a sense, although next to worthless medically.

Election-Night Forecasting

Each election night, when the polls have closed and the votes are being counted, the three television networks forecast the electoral outcome on the basis of early, partial returns—often needing only a few percent of the vote to predict accurately the final outcome. The networks invest millions of dollars in their electoral coverage, which allows their viewers to learn the results of the election several hours earlier than they might otherwise. Although this is perhaps a small yield for the investment, the scramble for early returns needed for the projection of the winner might, in some places in some elections, discourage corrupt election officials from greatly altering the real count of the vote—since the pressure of getting the vote count in may reduce the time needed to fix the returns.

For example, pressures for a timely count may curb such abuses as those in Illinois in the 1968 tabulation:

> For days before the election, the Chicago papers were full of tales of heavy crops of bums and derelicts being registered in West Side flophouses to provide the names for a fine Democratic turnout. And suspicion became certainty in the press rooms . . . when it was learned that "computer breakdowns" and "disputed vote counts" were holding the Illinois decision back. Veteran reporters could be heard explaining . . . how the game was played in Illinois: how both the iron Mayor and his Republican enemies downstate would "hold back" hundreds of precincts in an effort to finesse each other to give a hint of the size of the total they had to beat; how they would release a few precincts as bait to lure the other man into giving away some of his. . . .[7]

This suggests that the count of the vote is a rather unusual statistic. For most social and economic indicators, there is a tradeoff between timeliness and accuracy: the quicker we get the information, the greater the error. Sometimes the making of economic policy has been based on very short-run economic statistics—with a resulting reliance

[7] Lewis Chester, Godfrey Hodgson, and Bruce Page, *An American Melodrama: The Presidential Campaign of 1968* (New York: Viking, 1969), pp. 760–61.

on less accurate statistics—and more accurate figures might well have produced a different policy. In contrast to the usual case, however, a slow count of the vote often indicates vote fraud, or at least the opportunity for vote fraud.[8]

Although they may, in passing, reduce vote fraud, the central concern of the networks is to forecast the winner of the election (and, secondarily, the winner's share of the vote) on the basis of scattered and very incomplete returns. Two methods, both interpreting early returns with reference to a historical baseline drawn from previous elections, have been favored: (1) comparison of tonight's returns with the returns from previous elections at the same stage of the count and (2) comparison of tonight's returns from various counties with the returns from previous elections from those same counties.

The first method begins by constructing, on the basis of a previous election, a curve showing the relationship between the proportion of the vote reported and the proportion of the reported vote for the Democratic (or Republican) candidate. Figure 2-4 shows one such pattern, indicating that in this case a Democratic candidate who has more than about 40 percent of the vote when less than about 70 percent of the vote has reported can expect to win rather easily when all the returns are in. Such a pattern might result from the early reporting of certain Republican areas and a slower count in heavily Democratic areas. Thus the curve—called a "mu curve"—helps adjust for the bias favoring one party or the other in the sequence of early returns. Figure 2-5 indicates how this might be done. Tonight's returns are compared with the historical pattern of reporting, an appropriate adjustment for reporting bias is made, and the final projection is put on the air. In practice, the method is fancied up a bit—but still its basic defect persists: it relies on the assumption that the order in which the vote is reported remains the same from election to election. This assumption has led to several predictive disasters, and now mu curves only supplement other, more solidly based techniques.

One such predictive botch occurred during an election when a heavily Republican state first introduced voting machines. As a result, that state's flood of Republican ballots came in hours earlier than usual; the mu curve, believing that these were the same votes it saw in each election every four years, quickly projected a Republican landslide for president. Hours and hours later, John Kennedy won one of the closest presidential contests in history.

[8]The problem of inaccurate counts of the vote is not unimportant; political observers guess that two or three million votes are stolen, miscounted, or changed in a U.S. presidential election. Nobody has a good guess about the partisan advantage, if any, resulting from stolen votes. The advantage differs by state.

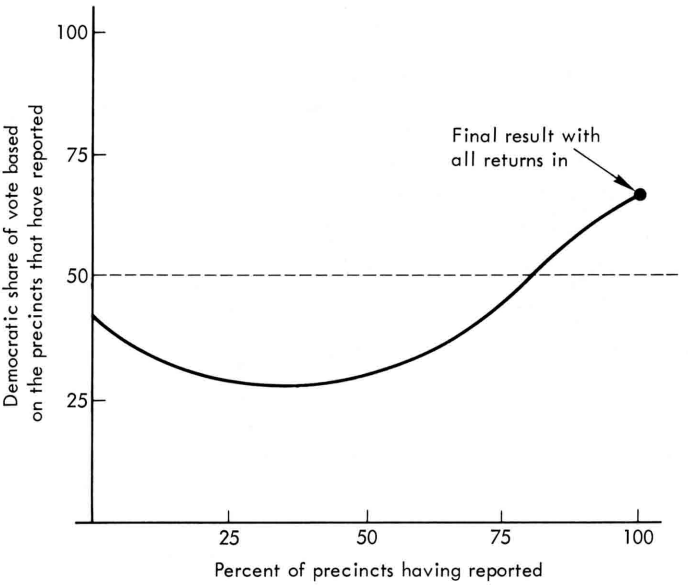


FIGURE 2-4 Historical pattern of the vote as more and more precincts report their returns on election night

Some practitioners patch up their mu curves by taking into account expected changes in the order of reporting:

> In deriving mu curves which are empirical in nature—they have to be—one must take into very careful consideration whether or not there have been any changes in voting patterns resulting from voting machines, or changes in poll closing times. Where there are such changes—and in every election we find that there are some—the mu curves have to be suitably adjusted in order to render them suitable.[9]

This sort of repair requires knowledge *in advance* of those changes in election procedures that might affect the sequence of the vote report—and must then guess how much earlier or later the affected returns will show up in the reporting sequence. The method also rests on the fragile hope that the patched-up curve traced out by tonight's returns will flow parallel to the historical curve—an assumption that will not hold up if there is a differential shift in particular

[9]Jack Moshman, "Mathematical and Computational Considerations of the Election Night Projection Program," paper presented at the Spring Joint Computer Conference in Atlantic City, N.J., on May 2, 1968, p. 3.

areas to a particular candidate. For example, if areas that normally report late and also normally vote somewhat Democratic suddenly shift very strongly toward the Democratic candidate because of that candidate's special appeal in those areas, then the paths traced out by the historical curve and tonight's curve would not be parallel, and the projection might be wrong. Finally, the method does not easily accommodate new political factors, such as a third-party candidate.

Because of these limitations and the availability of more powerful, more inferentially secure methods, mu curves are not now widely used in electoral projections, although they do retain some utility for informal use in interpreting election returns. That utility comes from the limited insight upon which mu curves are based: that different areas, with different voting patterns, report their returns at different times on election evening. Of course we knew that anyway.

The second—and preferred—forecasting method compares tonight's returns from those counties (or wards, precincts, or the like) that have reported early with the returns from previous elections in those same counties. The adjustment of current returns by previous per-

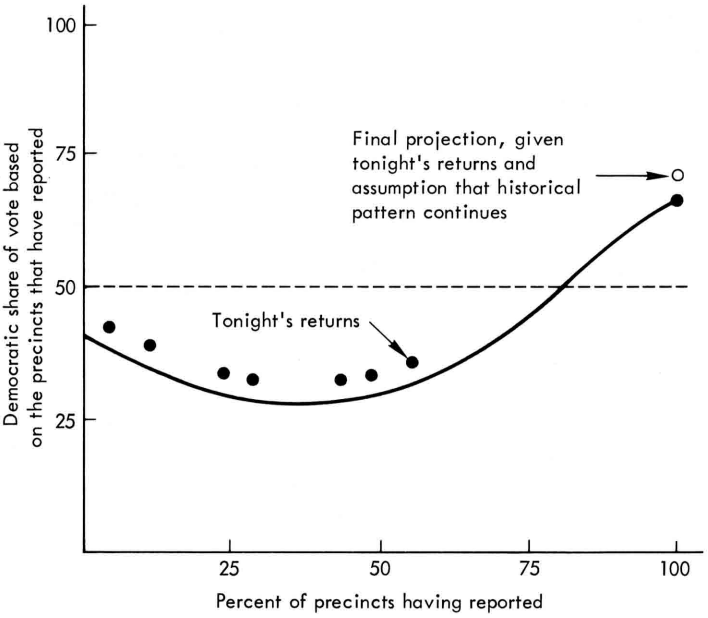


FIGURE 2-5 Comparing tonight's returns with the historical pattern to make a projection

formance at a disaggregated level (that is, at the county level) requires more detailed data and analysis than the mu-curve method—but it yields far more inferentially secure results. That is, there is a good chance that we know more after having done the analysis than we did before.

Comparing tonight's returns from a county with its previous voting patterns takes into account that the counties reporting first are not a representative sample. Counties with early complete returns may tend, in some states, to be Republican counties; in others, Democratic counties. At any rate, why hope they are typical or representative? Comparing current returns with old returns will adjust or control for a county's normal political leanings. For example, the raw returns from Massachusetts are not very helpful in projecting the national winner in a presidential race; but such returns are helpful if we know that Massachusetts normally runs heavily Democratic. So, if the Democratic candidate barely leads in Massachusetts, then that candidate is surely in real trouble nationwide.

Note the assumption here that the shift or the swing toward one party is roughly the same over the whole state or the whole nation. This assumption will not however lead to disaster—because it can be checked on election night with the data in hand simply by comparing the shifts across the counties that have reported. If the shifts are not consistent across counties, then either the historical base values from previous elections for the counties are ill-chosen and inappropriate for judging the pattern of tonight's election, or else the candidates had a special appeal to certain groups clustered by region and the shifts are not the same for different parts of the country. In contrast, violations of assumptions behind the mu-curve method are not easily discovered—at least in the short-run on election night.

Thus the second projection method is somewhat more powerful and safer than the use of mu curves because its assumptions are more modest and because some of its important assumptions can be verified during the course of the analysis. The second method does, however, require much more data and computing power; the grand assumptions of the mu curves are replaced by the collection and analysis of data.

In practice, the final projection of the election consists of a combination of several separate projections. This mixture forming the final, aggregate projection melds several component projections together:

1. the projection from the method of county-adjusted returns: $\%D_c$ = percent Democratic projected from counties;
2. the projection resulting from the so-called "key precincts," which are chosen either randomly or because of their special political

interest: $\%D_k$ = percent Democratic projected from key precincts;

3. the projection of the race before any returns are in at all, called a "prior"—a projection based on pre-election polls or political judgment: $\%D_p$ = prior projection of percent Democratic.

How much of each projection is mixed into the overall combined or "meld" projection? The prior, of course, receives full weight when no returns are in; as the returns pile up, the prior should carry less and less weight in the meld projection. Figure 2-6 shows one such weighting plan, with the weight, $w(r)$, a function of the number of precincts reporting. How should the other factors, $\%D_c$ and $\%D_k$, be weighted in the grand meld projection? Statisticians have a standard answer: form a weighted average using the reciprocal of the variances for weights.

Reciprocal weights are a reasonable choice—for, if the variance of an estimate is big, the weight should be small; if the variance of the estimate is small, then the estimate should have a relatively heavy weight and count for more because we have that estimate more precisely pinned down. Weighting by reciprocal variances gives, under ideal circumstances, the most precise combination. For the

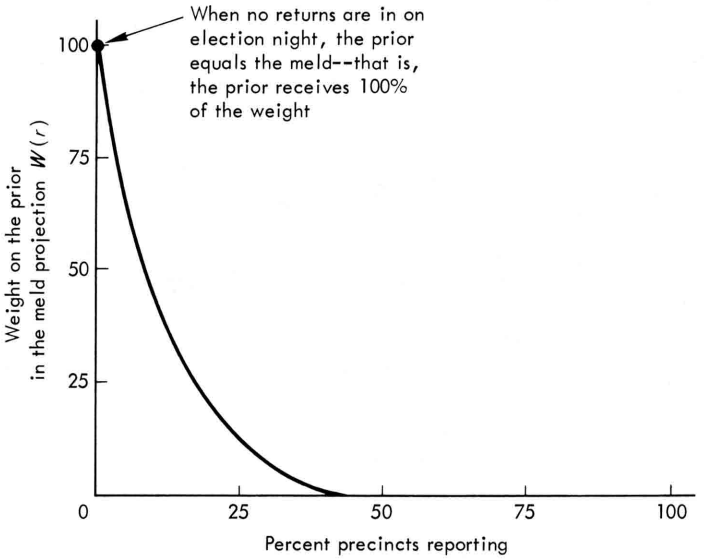


FIGURE 2-6 Weighting the prior in the overall meld

realities of election night, less simple combinations may be important. At any rate, one possible meld is the weighted average (weighted by the reciprocal of the variances) of the component projections:

$$\text{meld projection} = \frac{\frac{1}{S_c^2}\%D_c + \frac{1}{S_k^2}\%D_k + w(r)\,\%D_p}{\frac{1}{S_c^2} + \frac{1}{S_k^2} + w(r)}$$

where S_c^2 and S_k^2 are the variances of the estimates of $\%D_c$ and $\%D_k$. This is simply the particular realization of the general formula for a weighted average:

$$\text{weighted average} = \frac{\text{sum of weighted components}}{\text{sum of weights}} = \frac{\Sigma\, w_i x_i}{\Sigma\, w_i}.$$

Although based on the principles we have looked at here, contemporary projection models include many additional complications—complex estimation procedures, specially tailored base values, checks for bad data, and estimates of turnout. While today's elaborate models must be entirely computer based, in past years the votes were tabulated by hand on adding machines. Some years ago, the story has it, the truck delivering the dozens of rented adding machines to the studio on election day never arrived. Momentary panic arose, for how could they tabulate all the separate vote reports about to start pouring in? Finally, someone discovered a quickly available substitute for the adding machines. That night, ignoring the heavy-handed symbolism, they rang up the vote for president on cash registers!

Our next example evaluates another device for electoral forecasting—the "bellwether" district.

Bellwether Electoral Districts[10]

> Time present and time past
> Are both perhaps present in time future,
> And time future contained in time past.
>
> —T. S. Eliot, *Four Quartets**

[10]This section was co-authored with Richard A. Sun.

*From *Four Quartets* by T. S. Eliot. Reprinted by permission of the publishers, Harcourt Brace Jovanovich, Inc. and Faber and Faber Ltd.

Prior to the 1936 presidential election, the conventional political wisdom had it that as Maine voted, so went the rest of the nation. After the 46-state landslide, James Farley, Roosevelt's campaign manager, revised the theory: "As goes Maine, so goes Vermont." Such is perhaps the inevitable fate of so-called bellwether or barometric electoral districts; still, there are always new contenders with markedly unblemished records of retrospective accuracy to replace wayward bellwethers. Given the familiar inferential caution that retrospective accuracy provides little guarantee of prospective accuracy, what is the worth of claims that certain districts invariably reflect the national division of the vote?

The answers at hand differ: a skeptical statistician probably has little faith in the after-the-fact predictive success of bellwether districts; the collector of political folklore marvels at the record of such byways as Palo Alto County (Iowa) and Crook County, (Oregon) which have voted for the winner of every presidential election in this century; the newspaper reporter interviews a few citizens of Palo Alto or Crook County in search of "clues as to what will happen next Tuesday"; and Louis Bean has written four books premised on the notion that as goes X, so goes the country.[11] Here we will examine the question more deeply—and, at the same time, see a number of fundamental statistical techniques in action.

The data for the analysis are the election returns from almost all 3100 U.S. counties for the fourteen presidential elections from 1916 to 1968.[12] We will be looking for what are called "all-or-nothing" bellwethers: the county either votes for the winner of the presidential election or it does not. This seems to be the usual meaning of "bellwether district"; most discussions of supposed bellwethers report that the district has voted with the winner in the last N elections. Sometimes

[11] *Ballot Behavior* (Washington, D.C.: Public Affairs Press, 1940); *How to Predict Elections* (New York; Knopf, 1948) *How America Votes in Presidential Elections* (Metuchen, N.J.: Scarecrow Press, 1968); and *How to Predict the 1972 Election* (New York: Quadrangle, 1972).

[12] The data tapes were made available through the Inter-University Consortium for Political Research. We edited them extensively, correcting errors and adding missing data. Of the 3070 counties in the United States, we have the complete two-party election returns for the fourteen elections from 1916 to 1968 for 2938 counties, or 96 percent. The remaining counties had to be dropped because one election in the fourteen election series was missing; others may have changed names or are mixed in with other political units. A listing of the missing counties and election years was reviewed both before and after our analysis; both times it appears that the small amount of missing data had no consequences for our findings. Some of our early computations carried along votes for four different parties in each county, but we finally edited the data to include only the returns for the two major parties. Therefore all election returns reported here are based on the votes of the two major parties in all the elections.

N is surprisingly small; some journalists have interviewed nonrandomly selected citizens of "bellwether" communities that have voted for the winner in only three or four previous elections.

One good test of the credibility of bellwethers is to conduct a series of historical experiments, each designed to answer the question: How well would we have done in predicting the election of 19*XX* if we had followed a group of supposedly bellwether counties chosen on the basis of past elections before the election of 19*XX*? For example, going into the 1968 election, there were 49 counties that had voted for the winner in every presidential election since 1916—thirteen elections (or more) in a row with the winner. Were these 49 retrospective bellwethers more likely than other counties to support the winner in 1968? This is the sort of question that we will answer over and over, for different elections and for different choices of historical bellwethers.

Since they directly answer the question at hand, the historical experiments seem to provide the most powerful means of assessing the credibility of bellwethers. It is also possible to construct probability models to provide a baseline or null hypothesis against which to compare the observed performance of reputed bellwethers. We met with little success in developing models based on reasonable assumptions. The construction of a useful probability model remains an open question, although we suspect that even a very good model would still not provide as direct and powerful test of bellwethers as the historical experiment.

Another statistical problem arises because bellwethers are found in an after-the-fact search through election returns; there is no theory identifying particular areas as potential bellwethers before the fact. We have then a situation analogous to that of "shotgunning" in survey research: the searching through of a large body of data for statistically significant results leads to difficulties in just how to include the fact of the search in an adjusted significance test. One answer is simply the independent replication on a fresh collection of data of the results found through searching. That is, of course, the underlying logic of the historical experiment: bellwethers are chosen from a search, and then we see if their bellwether performance is replicated in the historical future.

The usual technique for evaluating bellwethers is retrospective admiration of the historical record. Almost all written accounts of reputed bellwethers describe an area's lengthy record in voting for winners and then ask, in effect, "Isn't that something?" These accounts evaluate the predictive performance of the past without reference to either prospective accuracy or the predictive record of other areas. Consider excerpts from a typical *New York Times* story on bellwethers:

> Town Votes 'Em As It Sees 'Em
> And It Usually Sees 'Em Right
>
> Salem, N.J., April 8—The political professionals are keeping an eye on this small Quaker community in southern New Jersey for clues to the outcome of the presidential election.
>
> For fifty years, with only two exceptions, Salem has voted for the victorious Presidential candidate. . . .
>
> There is no clear reason for Salem's stature as an election indicator.
>
> "But," says County Clerk Thomas J. Grieves, "you can't call it chance or a quirk. It happens too often. . . ."[13]

Actually, there are several hundred counties with predictive records better than Salem's over the last fifty years. But the important point is that no evaluation of Salem's record can be made on the basis of past election returns from Salem alone. A bellwether's credibility can only be assessed by examining, in comparison to other districts, its *predictive* record and not merely its postdictive record.

Consider the following historical experiment: let us choose the counties with the best records for predicting presidential elections from 1916 to 1964 and see how well they predicted the outcome of the 1968 election. There were 49 such counties with records of supporting the winner in all 13 elections from 1916 to 1964. Such a record, by almost any standard, is a bellwether performance—*if* the counties had been identified in 1916 instead of after the fact. How well did the 49 retrospective bellwethers of 1916–1964 do in predicting the winner in 1968? Not very well at all; 27 of the 49 (or 55.1 percent) voted with the winner in 1968. Two-thirds of *all* counties supported the winner in 1968, and so a county chosen at random could typically have been expected to outpredict the counties with previously perfect predictive records. Table 2-3 shows the full array of results, with the 1968 predictive performance tabulated against the prior record of predictive accuracy. Oddly enough, the best predictions in 1968 were made by counties that had had the worst record in the past (5 right, 8 wrong). These 80 counties (that went 100 percent for the winner in 1968) were, of course, counties that had voted without fail for the Republican candidate in every previous election since 1916 and persisted in 1968. So it is easy to find a group of counties, identified by their past voting record, that will support the upcoming winner—if you only know how the election is going to turn out!

The election of 1968 was a particularly bad year for the bellwethers of the past. Table 2-4, repeating the tests for the presidential elections from 1936 to 1964, shows that for some elections the bellwethers

[13] *The New York Times*, April 9, 1964, p. 29.

TABLE 2-3
Predictive Performance from 1916 to 1964 Compared with Predictive Record in 1968 Election

Past Performance, 1916–1964				1968 Performance			
Past Predictions		Counties		Right		Wrong	
Right	*Wrong*	*Number*	*Percent*	*Number*	*Percent*	*Number*	*Percent*
0	13	0	0.0	0	0.0	0	0.0
1	12	0	0.0	0	0.0	0	0.0
2	11	0	0.0	0	0.0	0	0.0
3	10	0	0.0	0	0.0	0	0.0
4	9	0	0.0	0	0.0	0	0.0
5	8	80	2.7	80	100.0	0	0.0
6	7	229	7.8	209	91.3	20	8.7
7	6	502	17.1	303	60.3	199	39.6
8	5	708	24.1	424	59.9	284	40.1
9	4	554	18.8	397	71.6	157	28.3
10	3	380	12.9	251	66.0	129	33.9
11	2	274	9.3	148	54.0	126	45.9
12	1	162	5.5	97	59.9	65	40.1
13	0	49	1.6	27	55.1	22	44.9
		2938	100.0	1936	65.9	1002	34.1

of the past do predict the upcoming election somewhat more accurately than a typical county.

Tables 2-3 and 2-4 provide us with a great deal of experience with retrospective all-or-nothing bellwethers. The tables suggest:

1. Perhaps each time one hears of an area with a spectacular predictive record in the past, a glimmer of hope and curiosity arises suggesting that surely this fine record couldn't be mere chance—there must be *something* going on. Whatever that something might be, it isn't a high degree of prospective accuracy. Sometimes previously accurate districts do better than just any collection of districts; sometimes they don't. The retrospective bellwethers were particularly poor in the close elections of 1960 and 1968. The compilations of Table 2-4 show the erratic record of the retrospective all-or-nothing bellwethers in predicting the future.

2. We have identified "bellwethers" in Tables 2-3 and 2-4 by their previously perfect predictive records in at least six consecutive previous elections. If this standard is applied to judging the results of our historical experiment, then the bellwethers of the past are not the bellwethers of the present. In five of the eight elections, the previously bellwether counties had a higher probability of voting with the winner

TABLE 2-4
Predictive Record of Previously Accurate Counties in Presidential Elections, 1940–1964

PREDICTING 1940	*Number of counties*	*Percent voting with winner, 1940*
1916–1936 past performance, right-wrong = 6-0	602	52.9
Nationwide	2938	61.6
PREDICTING 1944	***Number of counties***	***Percent voting with winner, 1944***
1916–1940 past performance, right-wrong = 7-0	319	72.7
Nationwide	2938	55.3
PREDICTING 1948	***Number of counties***	***Percent voting with winner, 1948***
1916–1944 past performance, right-wrong = 8-0	232	87.5
Nationwide	2938	59.9
PREDICTING 1952	***Number of counties***	***Percent voting with winner, 1952***
1916–1948 past performance, right-wrong = 9-0	203	81.3
Nationwide	2938	68.3
PREDICTING 1956	***Number of counties***	***Percent voting with winner, 1956***
1916–1952 past performance, right-wrong = 10-0	165	87.3
Nationwide	2938	70.0
PREDICTING 1960	***Number of counties***	***Percent voting with winner, 1960***
1916–1956 past performance, right-wrong = 11-0	144	35.4
Nationwide	2938	38.6
PREDICTING 1964	***Number of counties***	***Percent voting with winner, 1964***
1916–1960 past performance, right-wrong = 12-0	51	96.1
Nationwide	2938	73.3

than a county chosen at random from the nation as a whole; in the other three elections (1940, 1960, and 1968), a county chosen at random would be the county of choice in predicting the upcoming election.

3. The retrospective bellwethers, *taken as a group,* correctly predicted seven of the eight trial elections—in the sense that a majority of the group of retrospective bellwethers supported the winner. Exactly the same was true of a group of randomly selected counties (within the limits of sampling error).

4. There were, alas, no anti-bellwether counties. No county had such an outstandingly poor record that it could serve, by reversing its preferences, as a predictive (or even postdictive) guide.

5. Tables 2-3 and 2-4 indicate clearly why one obvious probability model, the binomial, for all-or-nothing bellwethers does not provide a useful baseline. Consider the following: if a fair coin, labeled "Democratic candidate will win" on one side and "Republican candidate will win" on the other, were tossed prior to each of the last 14 presidential elections, the probability that the coin would successfully predict the winner of all 14 contests is

$$\left(\frac{1}{2}\right)^{14} = \frac{1}{16{,}384} = .000061.$$

If this toss of the coin were performed in each of the 3100 counties, then it would be expected that

$$(.000061)\ (3100) = 0.2 \text{ counties}$$

would correctly go along with the winner 14 elections in a row. More generally, the binomial model for k successes in 14 independent trials with probability of success equal to one-half generates the distribution of predictions shown in Figure 2-7. The actual distribution of counties is also shown in the figure. It is clear that the distribution of actual election outcomes is not generated by a process of 14 independent trials with probability of success equal to one-half. That is because the probability of success usually substantially exceeds one-half and the trials are, in fact, highly dependent. The chances that a given county votes with the winner is usually around two-thirds, as Tables 2-3 and 2-4 show.

A more difficult problem in constructing a probability model is that the election results are not independent over space and time: both the interelection and intercounty correlations are very high. For example, the correlation between the division of the vote from one election to the next over all counties is almost always greater

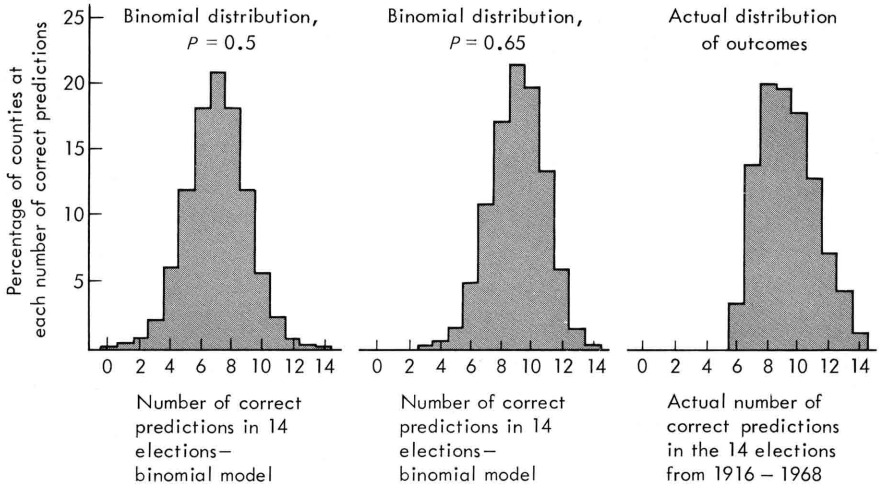


FIGURE 2-7 Binomial and actual outcome distributions

than .90. Considering that a county could go either Democratic or Republican in each of the 14 elections yields $2^{14} = 16{,}384$ theoretically possible electoral histories or paths that the counties could have followed over the 56 years. Less than 400 of these electoral histories actually occur, and only about 30 contain more than a handful of counties. At least 40 percent of all counties have gone more or less straight Democratic or straight Republican with occasional deviations in landslide years (Table 2-5).

TABLE 2-5
Most Frequently Occurring County Electoral Histories, 1916–1968

History	*Number of counties*
Straight Democratic	200
Democratic, except 1964	160
Democratic, except 1968	54
Democratic, except 1964 and 1968	58
Straight Republican	79
Republican, except 1964	128
Republican, except 1932, 1936, and 1964	136
Republican, except 1916, 1932, 1936, and 1964	155
Followed nation, all elections	27
Followed nation, except 1960	68

6. Twenty-seven of the nation's 3100 counties voted for the winner in every presidential election from 1916 to 1968. It may be possible—or at least a firm believer in bellwethers might well argue—that there are some truly bellwether districts hidden in those counties. What we have shown, of course, is only that counties with perfect postdictive records have undistinguished predictive records—when those counties are *taken as a group.* The only way we can identify bellwethers is as members of such a group. One final shred of evidence is to consider the performance of the nation's finest bellwethers. Prior to the 1960 election, there were eight counties in the nation with records of supporting every winner in this century. After 1968, only three of these eight superbellwethers still had unblemished records: Crook County, Oregon; Laramie County, Wyoming; and Palo Alto County, Iowa. They remained accurate in 1972.

Our conclusion in the case of all-or-nothing bellwethers is clear: the usual concept of a bellwether electoral district has no useful predictive properties. The all-or-nothing counties are only a curiosity and probably should be forgotten. It is a waste of time to send reporters out to interview nonrandomly selected citizens of Crook County a week or two before the election—at least it is a waste of time from any sort of scientific point of view. Such news reports create mystery where little exists.

There perhaps remains a magical air about the bellwethers of the past; some of these districts, considered individually, seemingly have such phenomenal records and yet we know better than to take them seriously—but still. . . . It may be best to look not to the election returns for the source of the mystery, but rather to ourselves. Maugham once wrote:

> The faculty for myth is innate in the human race. It seizes with avidity upon any incidents, surprising or mysterious, in the career of those who have distinguished themselves from their fellows, and invents a legend to which it then attaches a fanatical belief. It is the protest of romance against the commonplace of life.[14]

[14]Somerset Maugham, *The Moon and Sixpence* (Harmondsworth, Middlesex, England: Penguin Books, 1941), p. 7.

Regression Toward the Mean: How Prior Selection Affects the Measurement of Future Performance

Consider the defects in research design in the following example:

> Students in a statistics course who needed remedial teaching (as indicated by their performance in the lower quartile of an achievement test in arithmetic) were assigned to a special class in sensitivity training. Soon the teacher of the special class was able to go into full-time educational consulting because of the success of his new book, *Ending Educational Hangups in Statistics: How Empathy Pays Off*. The book showed that the special class was strikingly effective because when the students in the special class took the tests again after only six months, their test scores had greatly increased—increased, in fact, almost all the way up to the average of the first test scores of all the students who initially took the arithmetic test.

Several difficulties that are common in research designs compromise this hypothetical example.

This design uses the first test to divide the class into a treatment group (consisting of the lower quartile of students) and a control group (the remainder of the class). Students in the treatment group took the same tests again six months after joining the special class. The following comparisons were made in an effort to assess the benefits of the special class:

1. Average "gain" for special class equals

$$\left(\begin{array}{l}\text{average of scores on}\\ \text{second test for special}\\ \text{class}\end{array}\right) \text{ minus } \left(\begin{array}{l}\text{average of scores on}\\ \text{first test for special}\\ \text{class}\end{array}\right)$$

2. "Improvement" relative to rest of class equals

$$\left(\begin{array}{l}\text{average of scores on}\\ \text{second test for special}\\ \text{group}\end{array}\right) \text{ minus } \left(\begin{array}{l}\text{average of scores for}\\ \text{whole class on the first}\\ \text{test}\end{array}\right)$$

Two serious defects in the research design result in a bias in the "gain" and "improvement" scores such that the beneficial effect of the special class is exaggerated. The first defect is the failure to take into account the effect of practice and maturation on the test scores. Students taking a test a second time, as in the special class, can be generally expected to get better at taking tests; consequently, their scores improve merely because of their increased experience. Similarly, since the treatment-group scores on the second test are compared with the earlier test scores of the control group, a bias due to the maturation of the special group results. In other words, the students in the special group may improve relative to their previous performance (and the previous performance of their contemporaries) merely because they are older and smarter and not because they are necessarily benefiting from the special class.

In this design, then, the improvements in the scores of the special group due to practice and maturation effects are incorrectly attributed to the effect of the special class. Although it is impossible without additional information (or a better research design—see below) to judge the exact strength of the bias, we do at least know its direction: it favors the hypothesis that there is benefit from the special class.

The second defect in the research design is more subtle. It is a version of what is called the "regression fallacy." If members of a group are selected because their scores are extreme (either high or low) on a variable and if this extreme group are later tested once again, we will generally find that the group are "more average" than they were on the first test. Their scores will have moved or "regressed" toward the mean. One way to view the situation is to think of the extreme group as consisting of two sorts of people: (a) those who deserve really to be in that group and (b) those who are there because of random error—unlucky guesses on the test, an "off" day, and so forth. When the extreme group is tested a second time, the group (b) will typically perform more like their true selves, thereby raising their scores on the average at least. The deserving extremists in group (a) will continue their poor scores, albeit with some variation.

Thus the average score of the extreme group will typically increase because of the more typical performance of group (b) on the second test. There is no way of distinguishing group (a) from group (b) with only one test.

The problem arises when any group is formed by selecting its members because they are extreme on a single measure. For example, let us say that the highest quartile of students were placed in the special class instead of the bottom quartile. What would happen then? Once again, two types of students make up the extreme top group:

(a) those who are actually skilled and who deserve to be placed in the top quartile and (b) those who are lucky, who guess right, and so on. Now if this group is tested once again, it will generally be found that the overall average of the original extreme group has dropped somewhat—because not all the lucky performers on the first test will be lucky again.

The fallacy occurs in all sorts of situations. Wallis and Roberts provide several good examples including the following:

> Teachers—except, of course, statistics teachers—sometimes commit the regression fallacy in comparing grades on a final examination with those on a midterm examination. They find that their competent teaching has succeeded, on the average, in improving the performance of those who had seemed at midterm to be in precarious condition. This accomplishment naturally brings the teacher keen satisfaction, which is only partially dampened by the fact that the best students at midterm have done somewhat less on the final—an "obvious" indication of slackening off by these students due to overconfidence.[15]

Let us examine a numerical example of what might have happened in the case of the special class. Make the following assumptions:

1. There are no practice or maturation effects.
2. The special class has no effect at all on the students' test scores.

Under these assumptions we should observe no significant gains or improvements by the special class if the research design is free of bias. If, however, the research design has a bias, we will be able to get at least an approximate idea of its extent. Table 2-6 shows three sets of made-up test scores:

Column I: *The "true score" of each student on the test.* This, of course, is never actually measured perfectly, and the remaining columns represent the true score plus some random measurement error.

Column II: *The "true score" for each student with a random number between −20 and 20 added to each score.*

Column III: *Again the "true score" with another random number added to column I.*

Let the numbers in column II represent the scores of all the students on the first test and those in column III the scores on the second test. Since the test scores were computed by adding a random error to the "true scores," we find that there is very little difference in

[15] W. Allen Wallis and Harry V. Roberts, *Statistics: A New Approach* (New York: Free Press, 1956), p. 262.

TABLE 2-6
Random Errors Added to True Scores

	I		*II*		*III*
Student	*True score*	*Random error, test 1*	*Observed score, test 1*	*Random error, test 2*	*Observed score, test 2*
A	70	+13	83*	+1	71
B	75	−20	55*	+15	90
C	80	+8	88	−13	67
D	84	+7	91	−1	83
E	87	−15	72*	−9	78
F	90	+2	92	+8	98
G	93	−4	89	+12	105
H	95	−7	88	+16	111
I	96	+3	99	−12	84
J	97	+17	114	+20	117
K	98	−19	79*	−1	97
L	99	+11	110	+5	104
M	99	−18	81*	−17	82
N	100	−13	87*	+3	103
O	100	+9	109	−7	93
P	101	+12	113	+10	111
Q	101	−0	101	−5	96
R	102	−18	84*	+2	104
S	103	+13	116	+9	112
T	104	+7	111	−15	89
U	105	+3	108	+14	119
V	107	+12	119	−7	100
W	110	−11	99	+16	126
X	113	−20	93	+5	118
Y	116	+15	131	−19	97
Z	120	+1	121	+5	125
AA	125	−2	123	−2	123
BB	130	−14	116	−14	116

*The asterisk indicates students in lowest quartile on test 1.

the average score of the whole class on test 1 compared with test 2. Also the test seems to be measuring something: the correlation between the tests is .51. The correlation would be perfect, if we had not introduced the random measurement error into the true score on each test. Furthermore, note that the variability on both tests 1 and 2 is the same.

It should be clear that all that has been done is to construct some test scores containing some random error. No systematic effects in the data enable one to differentiate between the results of test 1 and test 2. But let us now see what happens in the research design

used in assessing the effects of the special class. The students in the special class were chosen because they were in the bottom of the class on the first test. Compare, then, the scores of the lowest seven students in the class as measured by test 1 (Table 2-7).

This research design generates the following misleading results. The average score of the group entering the special class was 77.3; after attending the special class for six months, their average score was 89.3—a "gain" of 12.0 points. Thus, because of the regression effects operating in this research design, a *pseudo-gain of 12 points* was found between test 1 and test 2, even though all the difference between test 1 and test 2 was generated by random numbers.

Note how plausible it all seems. A group of students are selected on the basis of test scores to enter the special class, and when the same students are tested later, those in the special class appear to have gained 12 points. Test 1 and test 2 are rather highly correlated, indicating that the tests are moderately reliable. And yet it is all a statistical artifact.

What would be a better research design—one that assesses the effect, if any, of the special class but avoids the bias resulting from the effects of practice, maturation, and regression toward the mean? The essential feature of an improved research designs is that not all of the low scorers should be placed in the special group. Ideally, some of the low scorers on test 1 should be randomly assigned to the special group; the others should remain in the regular class. In evaluating the effects of the special class, then, the basic comparison should be made between those low scorers in special class versus those low scorers in the regular class. Regression toward the mean still operates in this design, but its impact is roughly equal on the

TABLE 2-7
Scores on Test 1 Compared to Scores on Test 2 for the Lowest Quartile of Students on Test 1: Pseudo-Gains and Pseudo-Losses

Student	*Test 1*	*Test 2*	*Difference: "Gain" > 0 "Loss" < 0*
A	83	71	−12
B	55	90	35
E	72	78	6
K	79	97	18
M	81	82	1
N	87	103	13
R	84	104	20

control group and the treatment group because students were randomly assigned to the two groups.

The improved design, however, does give us a chance to separate out the genuine effects resulting from membership in the special class from the artifactual effects deriving from practice, maturation, and regression toward the mean. The original design confounds these factors and throws them all into the gain score.

This example also illustrates the utility of trying out the design and analysis on realistic but random data. Random data contain no substantive effects; thus if the analysis of the random data results in some sort of effect, then we know that the analysis is producing that spurious effect, and we must be on the lookout for such artifacts when the genuine data are analyzed.

Prediction of Accident Proneness: Can Producers of Automobile Accidents Be Identified in Advance as Consumers of Traffic Violations?

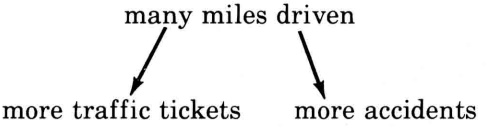
Only a small number of drivers are involved in severe automobile accidents. This fact gives rise to statements like "Three percent of all drivers produce one hundred percent of all severe accidents." The statement, while true, can be misleading. It does not mean that a small group of drivers go around systematically running down people or ramming other cars. "Accident proneness" may or may not be a useful concept.

It is empirically true that a small number of people, not necessarily identifiable in advance, are involved in serious accidents. Do these people have any characteristics in common? Can we ascertain roughly the probability that a given driver will be involved in an accident within a certain period of time? Insurance companies already make such predictions in a crude way by setting their rates in relation to factors including the driver's age, sex, marital status, accident history, type of driving, and record of traffic violations. Such procedures, at least as they are employed in Canada, are biased against some drivers (particularly high-risk drivers) because the various factors are not independent, resulting in double counting of risks against some drivers.[16]

[16]See R. A. Holmes, "Discriminatory Bias in Rates Charged by the Canadian Automobile Insurance Industry," *Journal of the American Statistical Association,* 65 (March 1970), 108–22.

A study of the relationship between the number of traffic violations a driver collects and his or her involvement in accidents is threatened by possible spurious correlations. First, one result of a motor vehicle accident is a traffic ticket. One driver or another is found to have committed a violation which "explains" the accident. This leads to statements such as "Accidents are caused by excessive speed," which are based on evidence that in many accidents, drivers involved are adjudged to have exceeded the speed limit. Lacking here is a comparison group of the speed of drivers *not* involved in accidents. There is some evidence that a large proportion of all drivers on the road are, in fact, exceeding the speed limit. In any case, a first step in a study of traffic violations and accidents is to control for the tickets produced by accidents—at least if the task is to predict, on the basis of a past history of traffic violations, that certain drivers will be more likely to be involved in accidents.

A second problem of potential spuriousness is suggested by the following model:



Thus, high-mileage drivers face greater exposure to the risk of both a traffic ticket and an accident—even if they drive with a care equal to that of low-mileage drivers.

A review of the studies of the relationship between violations and accident involvements points to both of these problems and to a partial solution:

> Ross investigated the relationship between violations and accidents for the 36 accident-involved drivers . . . and found that 12 of these 36 drivers had reported traffic convictions on their official records. These 12 people had 18 convictions. However, since there was no control group in this study, it is not possible to ascertain whether drivers with accidents had a higher violation rate than drivers without accidents. A point made by Ross, and one which has an important bearing on other studies using official records or information collected in interviews, is that there were discrepancies between interviewee-reported and recorded accidents and violations large enough to throw question upon studies relying on one or the other source of information in arriving at an accident or violation record.
>
> As part of a California driver record study, relationships between concurrent recorded accidents and citations (convictions for moving traffic violations) were analyzed. The data for this analysis consisted of a random sample of 225,000 out of approximately eleven million existing California driving records. Each driving record included a three-year history of both accidents and citations. To avoid inadvertent

correlation effects, citations directly resulting from accident investigations were labeled as "spurious" and were removed from the citation counts in most of the analysis.

The driver records were grouped according to the number of nonspurious citations, and the mean number of accidents per 100 drivers was calculated for each group. This analysis indicated an approximately linear relationship between citations and accidents with fluctuations at the high end of the citation count scale as a result of reduced sample size. Whereas those with no countable citations in the three-year period had only 14 accidents per 100 individuals, those with five citations had 62 accidents per 100 individuals and those with nine or more citations had 89 accidents per 100 individuals.

These figures indicate that there is a strong relationship between the mean number of accidents per driver and the number of concurrent citations when large groups of drivers are considered. On the other hand, the correlation coefficient between accidents and nonspurious citations was only 0.23. This low figure indicates that large errors could be made if one attempted to estimate the number of accidents an individual driver had on the basis of his citation record over the same time period. One would generally expect the correlation between concurrent events to be higher than nonconcurrent events. Thus, one should expect even larger errors, if one attempted to predict an individual's future accident record on the basis of his past citation record.

High-mileage drivers, other factors being equal, are exposed to a higher risk of both accidents and citations. Variations from driver to driver in exposure in general and annual mileage in particular may produce part of the correlation between accidents and citations that has been observed. Another California study examined characteristics of negligent drivers, defined as those whose record indicated a point count of four or more in 12 months, of six or more in 24 months, or eight or more in 36 months. (A point is scored for each traffic violation involving the unsafe operation of a motor vehicle or accident for which the operator is deemed responsible; two points are scored for a few types of violations deemed especially serious.)

When the annual mileage for a group of negligent drivers over age 20 was compared with that for a random sample of renewal applicants it was found that the negligent group averaged 17,219 miles per year while the applicant group averaged 7,449 miles per year. When males and females were treated separately it was found that negligent males averaged 17,591 miles per year as contrasted to 9,649 miles per year for the male applicants, while negligent females averaged 9,403 miles per year as contrasted to 5,519 miles per year for female applicants. The negligent drivers may have inflated their reported annual mileage in order to impress officials with their need to drive; nevertheless, it appears very likely that the negligent drivers do indeed drive more than average.[17]

[17] *The State of the Art of Traffic Safety,* by Arthur D. Little, Inc., for the Automobile Manufacturers Association, Inc. (Cambridge, Mass.: Arthur D. Little, Inc., June 1966) pp. 42–43.

Spellbinding Extrapolation

One of the most spellbinding efforts at simple extrapolation beyond the data arises in this history of guano:

> Guano, as most people understand, is imported from the [islands of the] Pacific—mostly of the Chincha group, off the coast of Peru, and under the dominion of that government.
>
> Its sale is made a monopoly, and the avails, to a great extent, go to pay the British holders of Peruvian Government bonds, giving them, to all intents and purposes, a lien upon the profits of a treasure intrinsically more valuable than the gold mines of California. There are deposits of this unsurpassed fertilizer, in some places, to the depth of sixty or seventy feet, and over large extents of surface. The guano fields are generally conceded to be the excrements of aquatic fowls, which live and nestle in great numbers around the islands. They seem designed by nature to rescue, at least in part, that untold amount of fertilizing material which every river and brooklet is rolling into the sea. The wash of alluvial soils, the floating refuse of the field and forest, and, above all, the wasted materials of great cities, are constantly being carried by the tidal currents out to sea. These, to a certain extent at least, go to nourish, directly or indirectly, submarine vegetable and animal life, which in turn goes to feed the birds, whose excrements in our day are brought away by the ship-load from the Chincha Islands.
>
> The bird is a beautifully arranged chemical laboratory, fitted up to perform a single operation, viz.: to take the fish as food, burn out the carbon by means of its respiratory functions, and deposit the remainder in the shape of an incomparable fertilizer. But how many ages have these depositions of seventy feet in thickness been accumulating!
>
> There are at the present day countless numbers of the birds resting upon the islands at night; but, according to Baron Humboldt, the excrements of the birds for the space of three centuries would not form a stratum over one-third of an inch in thickness. By an easy mathematical calculation, it will be seen, that at this rate of deposition, it would take seven thousand five hundred and sixty centuries, or seven hundred and fifty-six thousand years, to form the deepest guano bed. Such a calculation carries us back well on towards a former geological period, and proves one, and perhaps both, of two things—first, that in past ages, an infinitely greater number of these birds hovered over the islands; and secondly, that the material world existed at a period long anterior to its fitness as the abode of man. The length of man's existence is infinitesimal, compared with such a cycle of years; and the facts recorded on every leaf of the material universe ought, if it does not, to teach us humility. That a little

> bird, whose individual existence is as nothing, should, in its united action, produce the means of bringing back to an active fertility whole provinces of waste and barren lands, is one of a thousand facts to show how comparatively insignificant agencies in the economy of nature produce momentous results.[18]

Rather substantial inferences, given the observed data!

[18] *London Farmer's Magazine: Prospectus of the American Guano Company* (New York: John F. Trow, 1855).