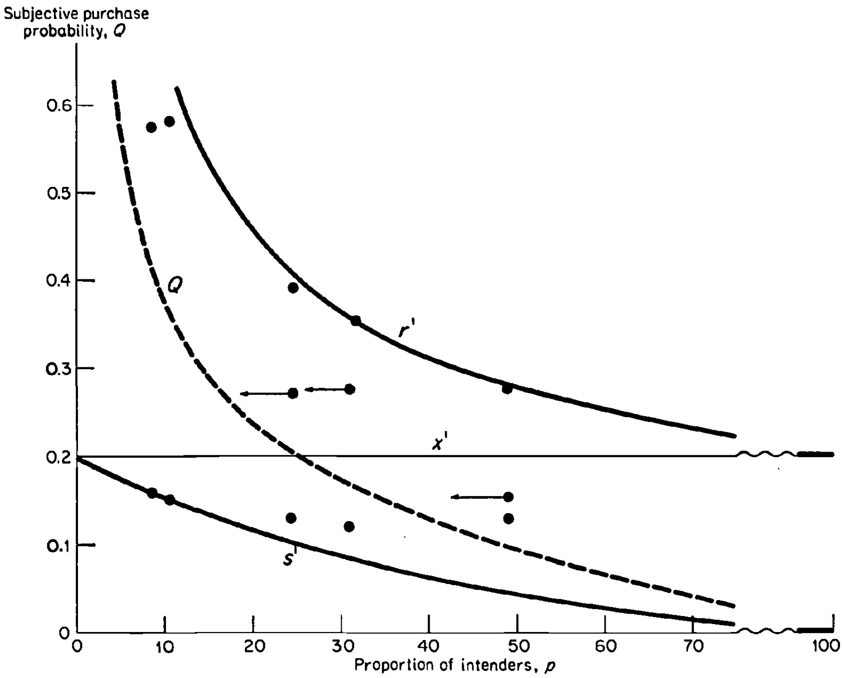
panel, where $x' = 0.5$, there is evidently a zero correlation between $r - s$ and p because the former is a constant (0.5) throughout the entire range of p. The last panel also has a distribution function with an x' of 0.5, but with the characteristics of a normal (Gaussian) function rather than of a rectangular function. In this panel there is zero linear correlation

CHART 8

Proportion of Sample Reporting Intentions to Buy Automobiles, and Proportion of Intenders and Nonintenders Purchasing Within a Six-Month Period



Source: Basic data from Tables 9 and 10.

between $r - s$ and p; however, there is a negative association between the two in the region where p is less than 0.5, a positive association where p is greater than 0.5. Finally, the variance in $r - s$, as a function of C or p,

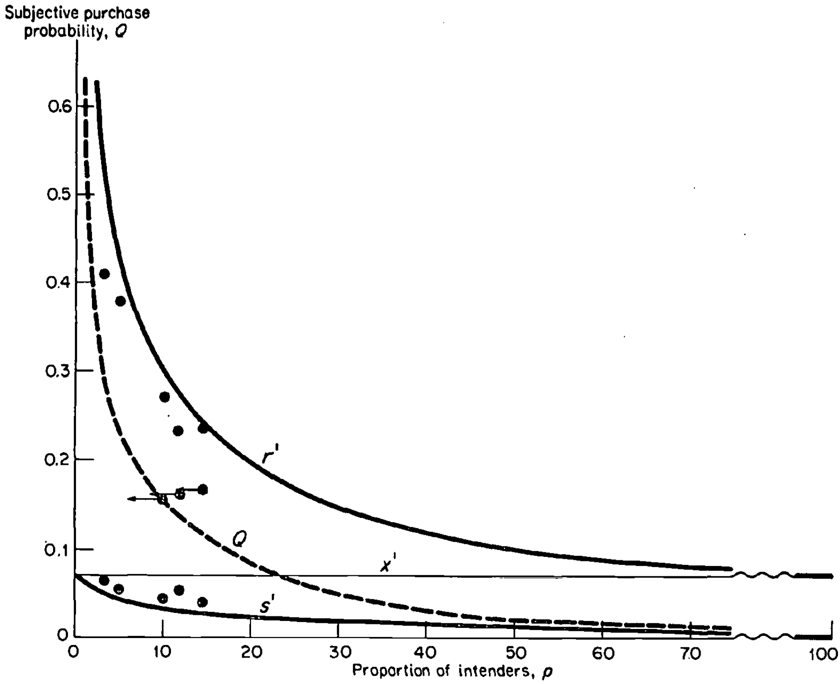
might well be positive for functions like those in the top panel; and for functions like those in the second panel, the correlation might well be negative for values of p close to zero, i.e., where nonintenders are close to 100 per cent. If functions were drawn for values of x slightly under 0.5 (top panel) or slightly over 0.5 (second panel) it would be impossible to tell by visual inspection whether the correlation between $r - s$ and p was positive or negative or zero in either panel.

depends on the value of x. The closer x comes to 0.5, the smaller the variance of $(r - s)$; the closer x comes to either 0 or 1, the larger the variance. All these generalizations can be observed in the data, as already discussed in the text.

It is an interesting exercise to plot some of the observed $(r,p)_i$ and $(s,1 - p)_i$ points to get a clearer view of the probability distributions for

CHART 9

Proportion of Sample Reporting Intentions to Buy Washing Machines, and Proportion of Intenders and Nonintenders Purchasing Within a Six-Month Period



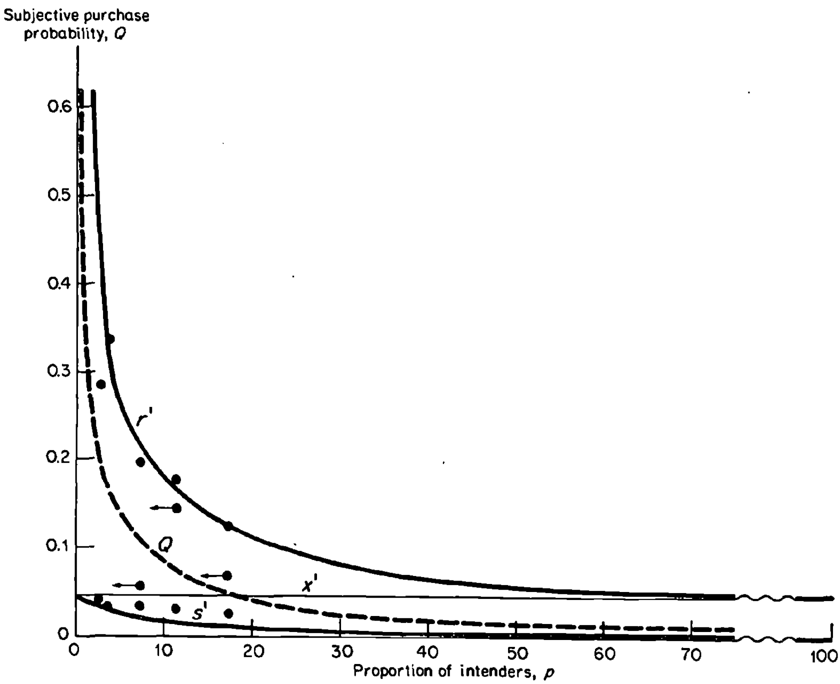
Source: Basic data from Tables 9 and 10.

different items. All observed functions have values of x that are below 0.5; consequently, functions like those in the top panel of Chart 7 are appropriate. The data for automobiles, for washing machines, and for clothes dryers are plotted in Charts 8, 9 and 10. These data are representative of the least skewed function (automobiles), the most skewed (clothes dryers), and the "typical" function (washing machines). Points on the r and s functions can be located precisely from the basic data in

Tables 9 and 10. Although some observations for the Q function are available, they all consist of averages for fairly large segments, that is, of mean probability between two unknown cut-off points. These are plotted at $(Q_{i \ldots j}, p_j)$; that is, the mean probability for households with probabilities lower than i but higher than j is located at p_j. Such points are necessarily too far to the right, and an arrow indicates this fact.[38]

CHART 10

Proportion of Sample Reporting Intentions to Buy Clothes Dryers, and Proportion of Intenders and Nonintenders Purchasing Within a Six-Month Period



Source: Basic data from Tables 9 and 10.

The empirically observed points in these charts are plotted from data in Tables 9 and 10. Panel A in these tables contains points on the Q function; Panel B has points on the r' function; Panel C, points on the s' function. The r' and s' functions are drawn to correspond roughly to the observed points, account being taken of the regression bias noted above. Because of this bias the observed $(r,p)_i$ points will generally be below the

[38] It cannot be correct to plot these q points midway between p_i and p_j, since the Q function is not a straight line. The $i, \ldots, j$ points are so far apart that location at the midpoint is a poor estimate.

true $(r',p)_i$ point, since *ex ante* probability must be higher than the *ex post* purchase rate for groups of households in the upper tail of the distribution. Conversely, the observed $(s,1-p)_i$ points will generally be above the true $(s',1-p)_i$ points. In Charts 8, 9, and 10 the bias is clearly evident in the s' function; any reasonable estimate of this curve must lie below the observed s points, especially for relatively high values of p.

Appendix II: An Estimate of the Potential for Improvement in the Prediction of Purchases

One additional piece of empirical data can be brought to bear on the prediction problem. Granting the probability hypothesis, there are only two reasons that a survey of buying intentions will produce an erroneous forecast of the population purchase rate: first, the survey may not yield enough information for a reliable estimate of mean *ex ante* purchase probability, because the proportion reporting that they intend to buy is a less than perfect substitute for the mean; second, the *ex post* purchase rate may diverge from *ex ante* mean purchase probability because of unforeseen events. Evidently, only the first type of error can be reduced or eliminated by improvements in survey design.[39] But are errors of the first type large enough to merit a serious attempt to reduce or eliminate them?

The question cannot be answered with precision, but an estimate of the minimum size of type 1 errors can be obtained from the data. In Chapter 2 estimates are given of the proportion of the variance in actual purchases explained by responses to each of the intentions questions—that is, the cross-section r^2 (coefficient of determination) between X and P, where both X and P are dichotomous (1,0) variables. But $r^2_{X.P}$ is less than unity because of the effects of the two types of error just discussed.[40] And an

[39] Errors due to unforeseen events can be reduced only if both the functional relation between *ex ante* probability and unforeseen events, as well as the events themselves, can be specified. In that case perfectly accurate contingent forecasts would be possible.

[40] It may be objected that the cross-section correlation would be less than unity even if a perfectly accurate measure of *ex ante* purchase probability were available for each household and no unforeseen events took place during the period. The observed values of X are inherently dichotomous—either a household purchases the commodity or it does not—but *ex ante* probability is presumably a continuous function. Thus, of 100 households with probabilities of 0.1, on the average 10 will buy and 90 will not buy; and the cross-section correlation r^2_{XQ}, where Q is purchase probability, will be less than unity.

The objection, while valid, does not affect the analysis. The observed values of r^2_{XP}, where both X (purchases) and P (intentions to buy) are (1,0) variables, are unbiased estimates of r^2_{xp}, where x and p are the proportions of purchasers and intenders drawn from a number of randomly selected subsamples. But if purchase probability were an available datum, the correlation between values of mean purchase probability and the proportion of purchasers drawn from the same randomly selected subsamples would be unity, assuming that no unforeseen events occur and that the subsamples are very large.

estimate of the maximum size of type 2 errors can be obtained from the data; hence, also, an estimate of the minimum size of type 1 errors.

The argument is as follows: errors due to the influence of unforeseen events essentially constitute a regression bias. As noted above in the section "The Problem of Bias," this is one of the reasons that r_i is a downwardly biased estimate of r'_i; s_i, an upwardly biased estimate of s'_i. Although there is no way to estimate the true value of r'_i, there are several ways of estimating the maximum value of r'_i, given the commodity and the buying intentions question.

To begin with, it has previously been demonstrated that the following inequality must hold by definition:

$$1 \geqq r', s' \geqq 0;$$

that is, both r' and s' must be between zero and unity. Further, both r' and s' are also constrained by the definitional relation

$$x' = r'p + s'(1 - p)$$

Since r' cannot be greater than unity nor s' less than zero, it follows that

$$\text{mâx. } r' = \frac{x'}{p}, \text{ and}$$

$$\text{mîn. } s' = \frac{x' - p}{1 - p}$$

Thus when p is greater than x', r' will have a maximum value of less than unity and s' a minimum value of zero; where p is less than x', r' will have a maximum value of unity and s' a minimum value greater than zero.[41]

The above analysis thus provides two independently derived upper limits to the value of r'_i, and two independently derived lower limits to the value of s'_i.[42] These are

$$r'_i \lesseqqgtr 1, \text{ and}$$

$$r'_i \lesseqqgtr \frac{x'}{p}; \text{ in addition}$$

$$s'_i \geqq 0$$

$$s'_i \geqq x' - p/1 - p$$

[41] Both the above constraints are ineffective when x is approximately equal to p. In such cases the value of (mâx. r' − mîn. s') will be approximately unity, and the value of mâx. $r^2_{X,P}$ will also be close to unity.

[42] Estimates of mâx. r'_i and mîn. s'_i can also be obtained from Tchebycheff's Inequality, but the procedure is more complicated and the results are much the same. Given the true mean and standard deviation of any distribution, Tchebycheff's Inequality states that no more than $1/k^2$ cases will be beyond k standard deviations from the mean, regardless of the shape of the distribution. Turning this theorem around, the pro-

An estimate of the maximum correlation between X and P can evidently be obtained from the above maximum or minimum estimates, since

$$\text{mâx. } r^2_{X,P} = (\text{mâx. } r'_i - \text{mîn. } s'_i)^2 \frac{p(1-p)}{x(1-x)}$$

These estimates (of mâx. $r^2_{X,P}$) constitute the proportion of variance in purchases that could conceivably be explained by buying intentions if unforeseen events are given the maximum possible weight; the residual variance (1 − mâx. $r^2_{X,P}$) must therefore be an estimate of the minimum unexplained variation in purchases attributable to the use of a dichotomous (1,0) buying-intentions variable, rather than a continuous distribution of *ex ante* probabilities, in making predictions. Table 18 summarizes these estimates for each of the thirteen commodities and for several of the buying-intentions questions; the observed proportion of the variance in X explained by P is also shown.

The data suggest that, although unforeseen events may have served to reduce the purchases-intentions correlation very considerably, any reasonable estimate of their importance still leaves a good deal of variation in purchases that cannot be explained by a dichotomous intentions variable. In very few cases does it appear that much more than two-thirds of the total variance in purchases could have been explained by intentions, making an extremely generous allowance for the part of total unexplained variance due to unforeseen events. I conclude that there is a considerable potential for improvement if a survey can be designed that will yield more information about the probability distribution. To what degree this potential can be realized is, of course, a question on which the available data shed no light at all.

---

portion of cases equal to $1/k^2$ can be no more than k standard deviations from the mean. Defining p as $1/k^2$, and assuming that the mean of the *ex ante* distribution, x', is equal to the observed purchase rate (x), it follows that p cases can be no more than the distance $x + \sigma'/\sqrt{p}$. But this is the same as saying that

$$\text{mâx. } C_i = x + \sigma'/\sqrt{p}$$

Given that $\sigma' = (r' - s')\sqrt{p(1-p)}$ (see above, p. 53), and that $s' = \dfrac{x' - pr'}{1-p}$ (see above, p. 52), and assuming that the distribution above C_i is rectangular, it turns out that

$$\text{mâx. } r'_i = \frac{1/2(1 + x' - x'\sqrt{1-p})}{1 - 1/2\sqrt{1-p}}.$$

Given mâx. r'_i, and mîn. s'_i, mâx. $r^2_{X,P}$ can be readily estimated. The distribution free assumption in the Tchebycheff theorem, however, typically provides estimates of mâx. r'_i that are much the same as those shown in Table 18.

TABLE 18

PROPORTION OF VARIANCE IN PURCHASES EXPLAINED BY ALTERNATIVE BUYING INTENTIONS QUESTIONS

Commodity	*Buying Intentions Questions*[a] A_1	C_1	D_1	$A_1 + A_2$
	OBSERVED PROPORTION (r^2_{XP})			
Automobiles	.080	.120	.079	.072
Furniture	.096	.118	.067	.063
Carpets & rugs	.077	.076	.027	.046
High-fidelity equipment	.043	.059	.060	.021
Range	.095	.130	.055	.079
Refrigerator	.098	.084	.089	.083
Washing machine	.055	.114	.068	.049
Television set	.010	.068	.040	.012
Air conditioner	.066	.072	.084	.082
Clothes dryer	.036	.063	.044	.050
Dishwasher	.103	.070	.078	.067
Food freezer	.053	.065	.047	.049
Garbage disposal	.044	.082	.044	.031
	ESTIMATED MAXIMUM PROPORTION (mâx. r^2_{XP})			
Automobiles	.373	.854	.753	.523
Furniture	.681	.580	.652	.487
Carpets & rugs	.740	.493	.504	.399
High-fidelity equipment	.642	.491	.500	.307
Range	.486	.735	.781	.598
Refrigerator	.432	.679	.892	.522
Washing machine	.421	.792	.685	.601
Television set	.459	.710	.606	.424
Air conditioner	.376	.725	.560	.427
Clothes dryer	.596	.551	.578	.374
Dishwasher	.725	.615	.610	.328
Food freezer	.584	.722	.506	.309
Garbage disposal	.647	.665	.553	.436

SOURCE: Data in upper panel derived from Chapter 2; data in lower panel estimated according to procedures described in the text.

[a] For a discussion of the intentions questions, see Chapter 2.

[b] The procedures for estimating mâx. r^2_{XP} are described in the text. An illustration is as follows: For question A_1—definite intentions to buy automobiles within twelve months—$x = 0.191$ and $p = 0.081$. Hence the appropriate estimate of mâx. r'_i is unity, since the alternative estimate—mâx. $r'_i = \frac{x}{p}$—yields a value of r'_i greater than unity. The appropriate estimate of mîn. s'_i is

$$\text{mîn. } s_i = x' - p/1 - p = 0.110/0.919 = 0.120.$$

Since

$$\text{mâx. } r^2_{XP} = (\text{mâx. } r'_i - \text{mîn. } s'_i)^2 p(1-p)/x(1-x),$$
$$\text{mâx. } r^2_{XP} = (1.0 - .120)^2 .081(.919)/.191(.809);$$
$$\text{mâx. } r^2_{XP} = .373.$$