FITTING AND ANALYZING POLYMODAL DISTRIBUTIONS OF CONSUMER CHOICES

Purnell H. Benson, Rutgers Graduate School of Business

The Problem of Describing Consumer Choices

This paper reports research experiences and insights from analyzing frequency distributions of consumer choices, culminating with the trimodal problem presented by data published by R. N. Reitter (1969). First, some propositions are offered as premises for work in analyzing distributions of consumer choices.

1. Most dispersions of consumer choices along variations in product characteristics are normal or moderately skewed. This conforms with the psychological fact that most distributions of psychological traits are normal or near-normal, arising as they do from probability principles of genetics or environmental chances. The distribution of educational or economic opportunity is markedly skewed. When it impinges upon human development, dispersions of human traits become skewed like that of taxable incomes, for which the leptokurtic lognormal distribution function can be used.

2. In consumer buying, the pressure of social convention and product availability mold consumer choices in ways which gives distributions stubby tails and sometimes no tails. The penalty of being different crowds consumer populations into platykurtic distributions, flattened with cut-off tails. For this situation, a distribution function frequently used in psychometric work is the arcsine, which is symmetrical without tails. What shall one do when one needs a platykurtic skewed distribution? The Beta distribution is one possibility, but lacks sufficient flexibility.

3. Often it does not matter what distribution pattern we assume consumer choices to have, if the measure of the product characteristic, which is the basis for consumer choice, is converted into normal deviate measurements. That is, sweetness of pudding can be measured with reference to the median as the zero point and the spreadoutness of choices by the consumer population as the scale unit.

By definition, when a product characteristic is measured in this way, a normal distribution will always fit perfectly. The true distribution could be rectangular, and the normal deviate values can still be assigned. What is wrong with this normalization? The difficulty arises when boundaries are drawn between consumer segments. It is usually assumed that these boundaries between product stimuli fall midway between them. If the true distribution is distorted, the midlines fall at different places each time a different method is used for stretching and contracting. This does not matter in market segmentation analysis if the boundary lines for consumer buying choices are very approximate. When the choices which consumers make depend very much upon habituation and advertising, the boundaries become zones, rather than lines.

4. It does matter how data are transformed if the market is highly segmented, as in hard goods or automobiles. Here the accurate representation of buying densities and segment boundaries is crucial. The more precise mathematical solution which suggests itself is to find the functional relationship between the normal deviate value for accumulated choices and the product characteristic physically measured.

One such relationship is a modification in the usual lognormal. Let the relationship between the normal deviate and the logarithm of the product characteristic be rewritten as an exponential relationship between the product characteristic and *e* raised to the normal deviate as its exponent and then multiply the whole exponential term by the normal deviate. The relationship obtained is skewed and platykurtic. Moreover, by making the exponent in the normal deviate a polynomial, the degree of kurtosis and the skewness are controllable.

In working with this modified lognormal, it can be cast in the form of linear regression if we estimate and subtract the median of the product characteristic Z, then divide the resulting value by the normal deviate X, and then take the logarithm of both sides of the expression (Benson, 1965).

$$\log((Z - A)/X) = B + CX + DX^2 + \ldots$$

One limitation is the poor behavior of the distribution in the vicinity of the tails. One or both wings ends in an upward projecting lip, instead of a tail, or else one of the tails may be greatly elongated. These are conditions which become pronounced when extreme skewness or platykurtosis bordering on rectangularity are sought. For any ordinary distribution of consumer choices, the tail problem is not serious. In extreme situations, the function becomes discontinuous at intermediate points as well as at the ends.

5. A contrast exists between the fitting of frequency functions by classical methods of statistics and what the consumer researcher is doing. (a) From a practical standpoint, he is not much interested in the tails of distributions, when he is describing distributions of buying choices. The tails mean much in hypothesis testing. But to find out where the best sales prospects are is not hypothesis testing in the usual sense. Hence the consumer researcher does not satisfactorily measure how well an empirical distribution is described simply by means of a chi square test of goodness of fit, which is much affected by the smaller cell frequencies in the tails. In consumer work, the correlation between actual and estimated sales densities is a more meaningful measure of goodness of fit of a distribution function.

(b) With some exceptions, irregular dispersions of buying choices do not arise from any of the familiar probability models. They depend upon product availability, a categorical matter, and social custom, a truncating matter. Exceptions include use of binomial and beta probabilities in analyzing media choice behavior of consumers (Benson, 1971; Greene and Stock, 1967). Also, if time of entrance of the buyer into the market place is a variable of consumer choice, then the exponential and gamma distributions are relevant, the first where a burst of promotion gets K percent of the remaining consumers to buy after each increment of time, and the second where a chain reaction takes place with each consumer telling others by word of mouth buy until all available consumers have bought.

6. Consumer researchers are restricted by lack of adequate data. This is first evident in the fewness of the points at which to measure frequencies of consumer choices. Consumers who try out products will usually try out one or two, and no more, except in the artificial situation of a food-testing laboratory. Moreover, hasty trials of many product stimuli are not typical of the leisurely situation in which consumers reach their buying decisions.

Nor are the data collected of one uniform kind. They may be multiple choice, paired, or monadic. For expensive hard goods, as TV screen size, the influence of advertising upon choice of screen size is much less and market data provide multiple choice data which are reasonably dependable.

More commonly, consumers are asked to try out products. One form of data is paired choices. From the standpoint of consumer judgement, these are incisive data. The problem to which Kuehn and Day (1962) addressed themselves is that two products which are much alike tend to get about a 50/50 split, whether the two products are near the middle of the distribution of consumers or at one side. If the dividing line between consumer choices puts 30% on one side and 70% on the other side, then two products on either side of the line should draw a 30-70 vote. They don't, and what can be done about it? The question has still not been fully answered as an operating thing.

Consumers can be asked whether their ideal buying choice is for a product to one side or the other of an experimental product tried out, that is, do they want more or less of a product characteristic than the experimental formulation given to them to try out? In effect, they are asked to position their choices relative to the product they look at. By taking the accumulated number of consumers who will buy a product up to the one tried, the cumulative frequencies needed for fitting distributions are obtained. Differencing the cumulative frequencies gives class frequencies.

Analysis of Reitter's Trimodal Distribution

R. N. Reitter (1969), following J. O. Eastlack (1964), felt he found an answer to lack of empirical data for multiple choices by asking consumers to mark on a line-interval scale where they feel their ideal choices to be. But whether they do this with precision leaves unanswered questions. As a questionnaire procedure, Reitter's method leaves something to be desired. However, he has published data exhibiting three modes, as shown in Figure 1.

The frequency bars have been smoothed, first by drawing lines which connect first and third quartile points within each class interval with those of the

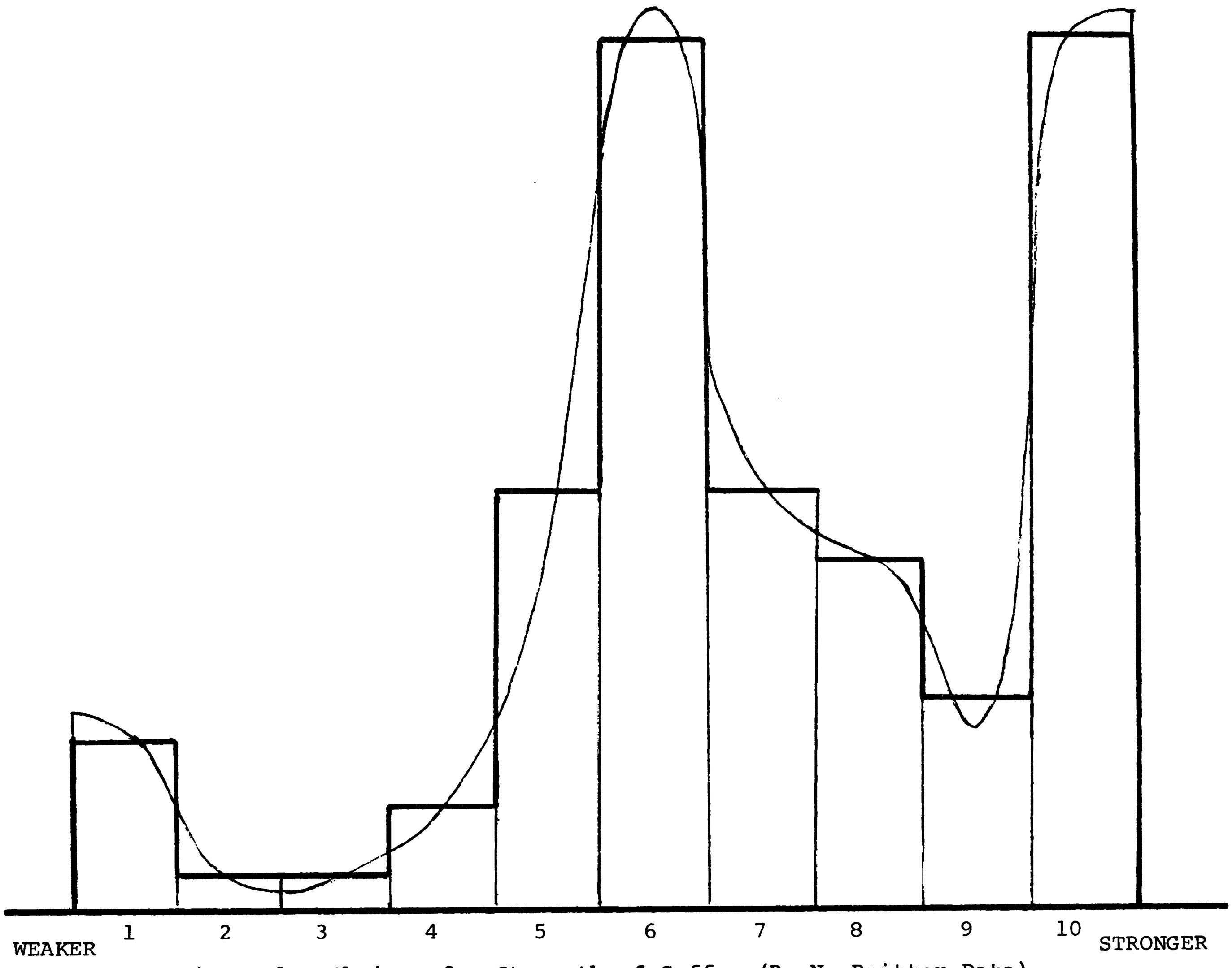

Figure 1. Choices for Strength of Coffee (R. N. Reitter Data)

next class interval, and then drawing a continuous curve which leaves areas between class boundaries unchanged. Coffee drinkers tend to check either the ends of this scale or the middle for how strong they want their coffee. Whether they are lethargic respondents or whether they really feel this way about coffee really does not matter if we are anxious to find data with which to explore the fitting of a polymodal distribution. The marks on the scale are choices made by consumers in some sense.

Before turning to the trimodal problem presented by Reitter, some comments may be made on fitting bimodal distributions. Such distributions are infrequent, and can usually be reduced to distributions of choices for related products, which have been combined. For example, lumping hand portable TV sets with room consoles will yield a bimodal distribution of buying choices. This bimodality can be handled by separating the data by product class to use unimodal analysis.

Otherwise, a workable procedure is to make a normal deviate conversion with the normal deviate a function of the product variable as a quartic polynomial or higher. In order not to waste the end points where the normal deviate is infinity, -3.0 and +3.0 are used as terminal values. Here only .002 of the distribution lie outside. If the data are not too erratic, this will yield a curve which appears decent on the surface. What is a decent curve? The statistician would say one which adequately fits properly collected data. The abundance of data for replicated tests of this kind is lacking. Making the accumulated area a function of the product variable, or vice-versa, is also serviceable to describe the distribution function. The modified lognormal distribution is so stable that it fills in the valley between the two modes. The more flexible linear polynomial is to be preferred.

Returning to the problem of fitting Reitter's data, it may be noted that the distribution has a central mode and J's at either end, and the left hand J reversed. The frequency distribution as it stands has no tails. One might infer that if more categories beyond the present ends had been offered, consumers might have checked them. In this event, the terminal frequencies represent cumulative frequencies up to the inside boundaries of the end categories. Here the class frequencies are taken as Reitter gives them and the distribution is regarded as ending with the outside boundaries of class intervals one and ten.

Eight functional types were tried out, as indicated in Table 1. Seven of them involve polynomial expansions on the right hand side. Terms up to the twelfth power were used, with best fitting terms selected by stepwise regression. This improves the fit for estimating class frequencies, but may do so at the expense of producing absurdities in the density function at intermediate points. Footnotes in Table 1 comment upon the behavior of the density functions.

Table 1

PERCENTAGE FREQUENCIES OF CONSUMER CHOICES ESTIMATED FOR TRIMODAL DATA, USING DIFFERENT TYPES OF FUNCTIONS FOR FITTING THE DISTRIBUTION

		Class Interval										
Type of Function[a]	Observed	1	2	3	4	5	6	7	8	9	10	Error Variance
	Frequency:[b]	5	1	1	3	12	25	12	10	6	25	
$Z = X + X^2 + ...$		3[c]	2	5	7	10	13	14	14	11	21[c]	24.6
$X = Z + Z^2 + ...$		5[c]	2	-1[d]	5	13	18	18	10	4	26[c]	10.0
$Z = Xe^{X+X^2+...}$		5[c]	1	1	3	16	21	12	9	8	24[c]	3.8
$X = Ze^{Z+Z^2+...}$		5[c]	4	5	8	11	13	13	11	10	20[c]	23.8
$A = Z + Z^2 + ...$		5	2	-1[d]	4	14	20	16	9	6	25	5.2
$Z = A + A^2 + ...$		5	1	2	2	13[e]	24	12	10	6	25	.4
$F = Z + Z^2 + ...$		4	-1	6	0	15	18	7	1	13	37	39.6
$A = \sin^2(X + pi/4) + ...$		4	2	1	7	14	17	16	14	8	17	18.6

[a] Z is measurement of product characteristic. X is horizontal deviate. A is accumulated frequency. F is class frequency.

[b] N equals approximately 200.

[c] These frequencies include accumulated proportions beyond class intervals 1 and 10. The density function rises abruptly to infinity outside of these class intervals.

[d] The density function is discontinuous within the class interval. Class frequency is obtained by differencing accumulated frequencies between boundaries of the class interval.

[e] The cumulative function is multivalued for this class interval. The frequency recorded is an average of two differences between class boundaries.

By coincidence, Reitter's data have a pattern which the modified exponential function could effectively represent. This distribution function has a central mode, and can terminate with upward projecting lips at either end. The fit of this function is included, along with simple polynomials which use both the product characteristic or the normal deviate as the dependent variable. Also tried are the accumulated frequencies as a function of the product characteristic, and also class frequencies as a function of the product characteristic. None of these functional formulations really copes satisfactorily with the data, as shown by Table 1. The modified lognormal is a fairly satisfactory fit. But left and right hand lips rise outside of the data.

The best fit, if the density problem is overlooked, is provided by the product characteristic expressed as a function of the accumulated choices. The most serviceable function from an all-round standpoint is the arcsine, listed last in the table.

This last line of solution regards the trimodal distribution as a composite of three separate distributions. Owing to the lack of tails, the use of the arcsine function for the separate modes seems appropriate. Each of the three component distributions has a mean, a standard deviation and a proportion within it. (The third proportion is dependent upon the first two, so there are eight unknowns). The 10 class intervals provide 9 degrees of freedom. There seems no simple solution to such a system other than taking trial values and generating a response surface which represents the residual error term in fitting the frequencies. The final result showed a component projecting to the left of the class with 6 percent in it and 1 percent adjacent. This last fit may imply that Reitter's data ought to include consumer replies before the first class interval and after the tenth class interval.

Still another method, not tested here, abandons the quest for a single mathematical expression to describe the density throughout the range of all of the frequencies. Instead, the frequency data are divided into several overlapping ones. Within each zone a separate distribution function can be fitted and densities calculated by taking the derivative of the cumulative function. This raises interesting questions concerning how well such overlapping functions will blend.

Some may feel that a good deal of time has been used collecting climbing equipment for a mountain climb still not completed. What does one do with a trimodal mountain so difficult to ascend? The problem is a broad one, arising in distributions of events over time, such as time series data, as well as in consumer choice analysis.

REFERENCES

Benson, P. H. Fitting and analyzing distribution curves of consumer choices. Journal of Advertising Research, 1965, 5, 28-34.

Benson, P. H. A solution to define latent segments of media audiences. 1971 Proceedings of the Social Statistics Section---American Statistical Association.

Eastlack, J. O. Consumer preference flavor factors in food product design. Journal of Marketing Research, 1964, 1, 38-42.

Greene, J. and Stock, J. S. Advertising Reach and Frequency in Magazines, New York: Readers Digest Association, 1967.

Kuehn, A. A. and Day, R. L. Strategy of product quality. Harvard Business Review, 1962, 40, 100-110.

Reitter, R. N. Product testing in segmented markets. Journal of Marketing Research, 1969, 6, 179-184.