PUBLICATION OF THE
GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
GEORGE F. BAKER FOUNDATION — HARVARD UNIVERSITY

VOLUME XXII MAY, 1935 NUMBER 3

THE USE OF STATISTICAL TECHNIQUES IN CERTAIN PROBLEMS OF MARKET RESEARCH

BY
THEODORE HENRY BROWN

DIVISION OF RESEARCH
BUSINESS RESEARCH STUDIES — NO. 12

PRICE $1.00

HARVARD UNIVERSITY
GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
BUREAU OF BUSINESS RESEARCH
SOLDIERS FIELD, BOSTON, MASSACHUSETTS

DIVISION OF RESEARCH: BUSINESS RESEARCH STUDIES

No. 1. Merchandising of Cotton Textiles — Methods and Organization, by Melvin T. Copeland and Edmund P. Learned $1.50
No. 2. Raw Material Prices and Business Conditions, by Melvin T. Copeland $1.00
No. 3. The Behavior of Consumption in Business Depression, by Arthur R. Tebbutt $1.00
No. 4. Department Leasing in Department Stores, by Stanley F. Teele $1.00
No. 5. International Raw Commodity Prices and the Devaluation of the Dollar, by Melvin T. Copeland .. $2.00
No. 6. Managing Cloth Inventories in the Cotton Textile Industry, by John J. Madigan $1.00
No. 7. Truck Selling: Simultaneous Selling and Delivery in Wholesale Food Distribution, by Lars J. Sandberg $1.00
No. 8. The Effect of Size on Corporate Earnings and Condition, by William Leonard Crum $1.00
No. 9. Management and the Worker: Technical vs. Social Organization in an Industrial Plant, by F. J. Roethlisberger and W. J. Dickson ..... 50 cents
No. 10. Corporate Earning Power in the Current Depression, by William Leonard Crum ..... 50 cents
No. 11. A Test of the Consumer Jury Method of Ranking Advertisements, by Neil H. Borden and Osgood S. Lovekin $1.00
No. 12. The Use of Statistical Techniques in Certain Problems of Market Research, by Theodore Henry Brown $1.00

BUREAU OF BUSINESS RESEARCH: BULLETINS IN PRINT

AUTOMOBILE TIRE AND ACCESSORY — RETAIL

No. 48. Operating Expenses in the Retail Automobile Tire and Accessory Business in 1923 ..... 50 cents

AUTOMOTIVE EQUIPMENT — WHOLESALE

No. 51. Operating Expenses in the Wholesale Automotive Equipment Business in 1924 ..... 50 cents
No. 42. Operating Expenses in the Wholesale Automotive Equipment Business in 1923 ..... 50 cents

BUILDING MATERIALS

No. 81. Operating Results and Policies of Building Material Dealers in 1928 ..... $2.50
No. 75. Operating Expenses of Building Material Dealers in 1927 ..... $1.50
No. 64. Operating Expenses of Building Material Dealers in 1926 ..... $1.50

CHAIN STORES

No. 95. Expenses and Profits of Variety Chains in 1933 ..... $1.00
No. 94. Chain Store Expenses and Profits: An Interim Report for 1932 ..... $1.00
No. 93. Expenses and Profits of Variety Chains in 1932 ..... $1.00
No. 90. Expenses and Profits of Department Store Chains and Department Store Ownership Groups in 1931 $1.00
No. 89. Expenses and Profits of Variety Chains in 1931 Compared with 1929 ..... $1.00
No. 87. Operating Results of Drug Chains in 1929 ..... $1.00
No. 86. Operating Results of Shoe Chains in 1929 ..... $1.00
No. 84. Expenses and Profits in the $1.00

CORDAGE

No. 82. Distribution of Hard Fibre $2.00

COTTON (*See also* TEXTILES)

No. 49. A Study of Cotton Hedging ..... 50 cents
No. 17. International Comparisons ... 19 — March, 1920 ..... 50 cents

DEPARTMENT STORES (*See also* C

No. 92. Operating Results of Depar $2.50*
No. 91. Operating Results of Depar $3.50*
No. 88. Operating Results of Depar $3.00*
No. 85. Operating Results of Depar $5.00*
No. 83. Operating Results of Depar $5.00*
No. 78. Operating Expenses of Dep ...ty Stores in 1928 ..... $2.50
No. 74. Operating Expenses of Dep ...y Stores in 1927 ..... $1.50
No. 63. Operating Expenses of Dep ... Stores in 1926 ..... $1.50
No. 61. Department Store Operatin ... 1923, 1924, 1925 — Preliminary Report $1.50
No. 59. Cases on Merchandise Con ...nent Stores (1926) $2.00
Operating Expenses in Department S ... 1921, No. 33 .. 50 cents each

45300
Brown:
Use of Statistical

*To firms furnishing figures for the departme ... Goods Association, which financed the work, and to educational institutions, etc., the list ... ns Nos. 88, 91, and 92 is $1.00 and the regular discounts apply.

Orders for the publications listed o ... of Business Research, Harvard Graduate School of Business Administrat ... henever possible the remittance should accompany the order. Checks sho ... Research.

Discounts: 50% to educational inst ... by universities, municipalities, governments, or public institutions. Trad

THE USE OF STATISTICAL TECHNIQUES IN CERTAIN PROBLEMS OF MARKET RESEARCH

BY
THEODORE HENRY BROWN, Ph.D.
Professor of Business Statistics

DIVISION OF RESEARCH
BUSINESS RESEARCH STUDIES—NO. 12

HARVARD UNIVERSITY
GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
BUREAU OF BUSINESS RESEARCH
SOLDIERS FIELD, BOSTON, MASSACHUSETTS

HARVARD UNIVERSITY

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION

GEORGE F. BAKER FOUNDATION

WALLACE B. DONHAM, *Dean*

MALCOLM P. MCNAIR, *Director of Research*

Publication of the Graduate School of Business Administration, Harvard University. Issued at Soldiers Field Station, Boston, Mass., eight times a year, during the months of March, April, May, June, July, August, October, and December. Entered as second class matter, May 17, 1913, at the postoffice at Boston, Mass., under the Act of August 24, 1912.

FOREWORD

The importance of the marketing problem since 1920 and the growing recognition of the dominant influence on marketing programs of the consumer's preferences, habits, desires, or even whims, as well as his capacity to purchase, have directed increasingly the attention of business to the need for market research. Large sums are spent annually by manufacturers, by distributors, by advertising agencies, and by marketing research organizations for the purpose of trying to learn more about the qualitative as well as the quantitative characteristics of markets for many different kinds of goods. Judged by the size of expenditures, market research itself is rapidly becoming a substantial industry. Mere size, however, is not the objective; for if the results are to justify the costs, market research has to be placed upon a scientific foundation which will include its own principles, special techniques, and procedures. It is toward the development of such special principles, techniques, and procedures that this study is directed.

Only in exceptional circumstances is it possible to canvass an entire market for consumer goods. Consequently market investigations usually proceed on a sampling basis. One of the first questions which the research director of a manufacturing company or an advertising agency must seek to answer is, "How large must the sample be?" To judge from published material on this subject, there are numerous opinions as to the size of sample necessary under various conditions. Usually conclusions in regard to the necessary size of sample seem to be reached wholly on the basis of the total number of all possible cases, which, in statistical terminology, is the size of the universe. Until recently, little attention has been given to an even more important factor, namely, the degree of accuracy required in the answers to the questions under investigation. This matter of size of sample is related also to a second group of questions that must be answered in the interpretation of differences observed in two or more percentages obtained from the sample. Quite customarily, percentages which have nearly equal numerical values, such as 47% preferring product A and 53% preferring product B, are regarded as not significant, with the result that attention is given principally to those percentages which show wide differences. In some cases the conditions of the problem may require differentiation between percentages as close as those cited; and, by the use of suitable techniques, questions as to the significance of such percentages are susceptible of much more exact determination. There are several related problems in this second group: the problem of error in a single percentage, the problem of error in several percentages involving independent unlimited choice, the problem of error in several percentages involving limited choice, and the problem of averages from samples.

Although these problems in point of time are subsequent to the initial problem of determining the necessary size of the sample, yet logically they afford the easiest

method of approach. The object of this study, therefore, is to present and explain the statistical techniques suitable for dealing with these related problems of the interpretation of percentages and the size of sample.

The illustrative data which have been used were obtained through the generosity of several leading advertising agencies. In each case, titles of tables of data have been changed or the data adjusted so that no confidential information is revealed. The reader naturally is to assume that the data presented in this study have no importance other than that of an illustrative character.

It is important that the reader keep in mind the practical limitations of the procedures described in this bulletin. Questions of the randomness of the sample and of the consistency and accuracy of the data may be far more important in individual cases than questions of interpretation of differences in percentages. The techniques described in this bulletin are not to be used in the naive belief that the results will be accurate in proportion to the number of decimal places to which the calculations are carried. Yet if these procedures are used with a proper appreciation of their limitations they should afford a helpful tool for market investigators to use in seeking to eliminate a portion of the guesswork sometimes present in the interpretation of market data.

This report by no means undertakes to cover all the statistical techniques which are appropriate to the problems of market investigation, or even all those which relate to the problems of sampling. Continuing research is expected to result in publication at a later time of additional studies relating to the statistical procedures applicable to many of the other problems encountered in the task of market analysis. This part of the study is being published at the present time partly in response to the requests of several organizations which wish to make immediate use of these techniques, and partly for the purpose of facilitating a wider test of the practical applicability of these procedures.

In the preparation of the bulletin the valuable assistance of Mr. Dickson H. Leavens, Miss Grace E. Crockett, and Miss Frances V. Scott is gratefully acknowledged.

THEODORE H. BROWN

MAY, 1935

THE USE OF STATISTICAL TECHNIQUES IN CERTAIN PROBLEMS OF MARKET RESEARCH

SECTION A. SAMPLING ERRORS

PROBLEM 1. SINGLE PERCENTAGE

The solution of a problem involving the interpretation of the difference between two percentages is attacked most easily by determining first the possible error in a single percentage obtained through a sampling process. This single reported figure in a market analysis usually expresses in percentage form the ratio of the number of favorable replies to the total number of replies included in the sample.

In gathering the data, it is assumed that a random[1] selection is made from the market which is being examined. Anyone will concede that a second sample gathered under such conditions would be likely to produce a number of favorable replies which would give a percentage somewhat different from that obtained in the first sample. Moreover, it is reasonable to expect that with a large number of samples there will be a concentration about the true percentage with larger and larger deviations from this true value occurring less and less frequently.

Thus, if we knew the true percentage value for the whole universe (market), the chance that a given sample would have a deviation of a given size could be estimated. Unfortunately, in practical problems the true percentage value is not known.[2] Consequently, by a converse reasoning, the percentage as given by the sample is adopted as the standard, and an estimate is made of the limits within which probably the true value would occur if a complete census of the universe were possible.

The estimates of limits within which the true value probably lies is based upon the theory of probability or chance. The measure which is used for such estimates is known as the standard deviation.[3] On the basis of certain theoretical calculations and practical experience it has been found that three standard deviations on either side of the observed value probably will include the true number of favorable replies. In the use of such a measure, the chance of being wrong seems to be not more than 1 in 100 times. This chance of being wrong not more than 1 in 100 times is accepted as a practical criterion for the limits within which the true value will occur.

Application — Parker Company[4]

The Parker Company, which manufactured toothpaste, had engaged an agency to investigate the market for its product. Because of the expense involved in a census, the agency planned to base its conclusions on a sample which would represent the character of consumer demand for toothpaste. In the report submitted to the company, the agency included a table which indicated the number of users of paste in comparison with the number of users of other types of dentifrice, such as liquid or powder. These data were as follows:

Table 1. Preference for Toothpaste

Users of Toothpaste	2,768	69.2%
Users of Other Forms of Dentifrice	1,232	30.8%
Total Replies to Inquiry	4,000	100.0%

From a study of these data it seemed to the research director of the Parker Company that the users of paste represented a decided majority of the consuming public. He recognized, however, that another sample might give quite different results. Doubt as to the certainty of the interpretation raised the question whether there might not be some way of measuring the error in the percentage of the users of paste as indicated by the sampling results shown in Table 1. It was this error in the percentage which the research director of the Parker Company desired to know.

Technical Approach

A study of this relatively simple table of data indicates that there are three factors present whose influ-

[1] For a discussion of random samples, see page 11.
[2] For a discussion of what is known as *a posteriori* probability, see page 16.
[3] For an interpretation of the meaning of standard deviation, see page 16.
[4] Fictitious name.

ence may be used to determine the size of the error. These are the percentage of users of paste, the percentage representing those who use some other form of dentifrice or none, and the number of individuals who are questioned.

These same three items appear so frequently in problems of this type, and others which will be developed, that specific letters are used to indicate them. Thus, the proportion of favorable occurrences, which in this case is the number of individuals using paste, is denoted by the letter p. Similarly, the proportion of unfavorable occurrences, which here represents those using other forms of dentifrice, is indicated by the letter q. It is at once obvious that $p+q=1$. In this particular equation it is assumed that p and q will be expressed in decimal form and not in the form of percentages. Thus, for the given table

$$p=0.692,$$
$$q=0.308.$$

Hence, in decimal form

$$0.692+0.308=1.000.$$

The values of p and q may be expressed in terms of chance of success or failure. Thus, if one of the 4,000 people interviewed for this problem be selected at random, the chance that that person will use paste is 69 out of 100 or 0.69 out of 1. Similarly, q represents the chance of failure. Consequently, the equation states that the chance of success plus the chance of failure is certainty.

Finally, the number of people interviewed, which in this case is 4,000, is denoted by the letter n.

On the basis of the three facts indicated by the letters p, q, n, it is necessary to build the measure of the error involved in the sampling process. Since such a measure also is used frequently, a standard letter is employed. This is the standard deviation, σ. Mathematicians have derived an expression for this measure of error in terms of the three basic letters.[1] The formula is

$$\sigma=\sqrt{npq}.$$

In many situations the results of the investigation are given in terms of a percentage of the total. The symbol $\bar{\sigma}$ is used for these percentages. For such cases[1]

$$\bar{\sigma}=\sqrt{\frac{pq}{n}}.$$

As already indicated, three standard deviations are used as a measure of the limits involved.

Solution

Since the data in Table 1 are converted into percentages of the total, we choose for estimating the error the formula

$$\bar{\sigma}=\sqrt{\frac{pq}{n}}.$$

Because p represents the chance that a case selected at random will be favorable, the value of p in this case is 69.2%. Similarly, as q represents the chance of failure, its value is 30.8%. Both these values are based upon the 4,000 inquiries. Consequently, n equals 4,000. The table of values and the calculation may be summarized as follows:

$$p=69.2,$$
$$q=30.8,$$
$$n=4,000,$$
$$\bar{\sigma}=\sqrt{\frac{69.2 \times 30.8}{4,000}}$$
$$=0.7\%.$$

As indicated above, in order to exclude all but the extreme 100 to 1 chance, 3σ is used as the error. The limits for the problem thus become $3\bar{\sigma}=2.1\%$. Hence the investigation shows that the percentage of those who prefer toothpaste equals $69.2\%\pm2.1\%$ of the population. These values presumably will not be exceeded more than 1 in 100 trials of 4,000 cases each.

Conclusion

The calculated error of 2.1% indicates, on the basis of the above reasoning, that the users of paste comprise two-thirds of those consumers using any type of dentifrice. Obviously, the research director was quite right in his opinion that the proportion of those using paste was substantial in relation to the size of the sample. On the other hand, knowledge of the error involved should clear up any possible doubt in his mind as to the significance of the result. To illustrate, assume that the standard deviation had proved to be equal to 7%, the value which would have been found if a sample of 40 had been taken instead of 4,000. Then three standard deviations would be equal to 21%. Subtracting this amount from the observed value of p, the true value obviously might have been as low as 50% so that q would have been approximately 50%, with the odds of 100 to 1 against coming outside these limits. For such a case the indicated consumer demand for paste would be equal to the indicated consumer demand for other types of dentifrice.

The importance of the result obtained in this case, however, does not lie wholly in the solution of this single problem. The method of attack, as well as the solution procedure, is of value in more complicated problems.

[1] For proof of the formula, see page 18ff.

PROBLEM 2. SEVERAL PERCENTAGES: UNLIMITED CHOICE

A frequent problem in market investigations involves the decision as to whether an observed difference between two percentages is or is not significant. Thus, if the market preference for Brand A is indicated by 30% of all the votes cast, while the preference for Brand B is indicated by 25%, it is necessary to decide whether another sample might not reverse these two brand preferences.

There is an additional point which often causes trouble. This is the question whether the choice granted to the person being interviewed is in effect unlimited or limited. Usually the decision as to whether the data should be analyzed in accordance with methods for unlimited or limited tables of choices depends upon the form in which the question is phrased. Thus, if Mrs. Jones is asked to name her preference of any one of four department stores in her city, the choice essentially is limited. If, on the other hand, Mrs. Jones is asked to name the brand of toilet soap which she prefers, for practical purposes the number of choices which she may make is unlimited. It is recognized that the method of solution suggested for such cases is not mathematically correct, but the practical exigencies of the problem permit the analysis as if the number of opportunities from which Mrs. Jones might make her choice is unlimited.

The first step in the solution of the problem is to find the error in each percentage just as has been done in Problem 1. The single added element which distinguishes this problem from Problem 1 then will be the fact that the difference between the two percentages must be considered. Consequently a formula extending our information involving this added factor must be used.

Application — Benson Company[1]

The Benson Company had been retained by a manufacturer of toilet soap to make a market analysis. Among other data gathered from a sample of 534 housewives were the figures shown in Table 2, indicating preferences for various brands of soap.

The research director of the Benson Company received the report, but he was somewhat puzzled as to how to interpret the percentages in Table 2. For many of the brands the opinions of the housewives were divided so evenly that no clear preference was apparent. Moreover, in other parts of the table the number of votes indicating preferences graded from one brand into the next so gradually that the research director found himself quite perplexed as to how to make a proper interpretation. The problem which the research director faced was that of deciding whether the observed difference in choice between any two brands, such as Brand A and Brand B, might not be eliminated or reversed in another sample.

Table 2. Preferences for Brands of Toilet Soap

Brand	Number	Per Cent
A	60	11.2
B	56	10.5
C	40	7.5
D	34	6.4
E	24	4.5
F	22	4.1
G	20	3.7
H	14	2.6
I	14	2.6
J	12	2.2

Percentages based on 534 answers to question.

Technical Approach

As discussed in Problem 1, the errors in the individual brand percentages can be calculated. When it is necessary to decide whether the observed difference in percentages between two brands is significant, a somewhat more complicated formula is needed. It is assumed that the percentages for each of two brands have been obtained from the data. In addition, it is assumed that the percentage error for each of the two percentages has been calculated. Presumably a true value for each percentage would lie within the error limits in each case. Fundamentally, then, the question becomes one of deciding whether these two true values might not be equal. If this were the fact, the common true percentage would lie between the two observed percentages. So far as the separate brands are concerned, such a condition would exist when the true percentage occurred within the positive permitted error of the lower of the two brand percentages and within the negative error of the larger of the two. Clearly, it would hardly be fair to

[1] Fictitious name.

permit ourselves to use the extreme values of each case, adding them in order to obtain the comparison error. The reason is that this would be more improbable than the probable standard of $3\bar{\bar{\sigma}}$ which we have set as our limit for each case. Consequently some method of combining the individual values to give a somewhat smaller permissible error is necessary.

In this case, as in Problem 1, a mathematical analysis shows that if the two percentages, and consequently the choices as made in the market investigation, are entirely independent of each other, the relationship that is sought may be expressed in terms of the standard deviations.[1] This is

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2.$$

In the above formula

σ_u is the standard deviation of the difference u between the two percentages x and y,

σ_x is the standard deviation of the percentage x,

σ_y is the standard deviation of the percentage y.

In percentage form the formula becomes

$$\bar{\bar{\sigma}}_u^2 = \bar{\bar{\sigma}}_x^2 + \bar{\bar{\sigma}}_y^2.$$

Solution

The question raised in regard to the observed difference between Brands A and B is whether the difference of 0.7% between 11.2% and 10.5% is important. Could another sample remove the apparent difference in percentages or even reverse the brand ranking? For this problem the data available for each brand are similar to those given in the case of Problem 1.

The steps involved in the solution, consequently, are as follows:

Brand A	*Brand B*
$p = 11.2$	$p = 10.5$
$q = 88.8$	$q = 89.5$
$n = 534$	$n = 534$
$\bar{\bar{\sigma}}_x^2 = \frac{11.2 \times 88.8}{534}$	$\bar{\bar{\sigma}}_y^2 = \frac{10.5 \times 89.5}{534}$
$\bar{\bar{\sigma}}_x^2 = 1.86$	$\bar{\bar{\sigma}}_y^2 = 1.76$

Hence,

$$\bar{\bar{\sigma}}_u^2 = 1.86 + 1.76,$$
$$\bar{\bar{\sigma}}_u^2 = 3.62,$$
$$\bar{\bar{\sigma}}_u = 1.9.$$

The observed difference between Brand A and Brand B is 0.7%. Obviously, $3\bar{\bar{\sigma}}_u$ or 5.7% is far larger than the observed difference. Consequently, it cannot be assumed that there is any difference in market demand between Brand A and Brand B.

From the data as developed it is possible to make some preliminary guesses as to the point in the table at which the differences in the data become significant. The yardstick $3\bar{\bar{\sigma}}_u$ as determined from the calculations is nearly 6%. Consequently it would be expected that the difference between Brand A and Brand D or Brand E would be sufficiently great so that it would not be eliminated if another sample were employed.

If the calculations are completed for the significant differences between Brand A and Brands C, D, and E in turn, the following results are obtained, including the test for Brand B already worked out in detail:

Original Data		Observed Difference from Brand A	Permissible Error in Difference $= 3\bar{\bar{\sigma}}_u$	Interpretation
Brand	Preference in Per Cent			
A	11.2	—	—	—
B	10.5	0.7%	5.7%	Not Significant
C	7.5	3.7	5.3	Not Significant
D	6.4	4.8	5.1	Doubtful
E	4.5	6.7	4.9	Significant

Thus the range of sampling errors is so large for these data that we cannot say, except for widely separated brands, that the differences in the values are significant.

Conclusion

In this problem it is expected that the preferences for each brand will be different because it is tacitly assumed that the market demand is different for each brand. The results of the analysis, however, indicate that the first three brands, A, B, C, have no respective preferential rating and that possibly D might be added as a doubtful fourth to the list. It may be, of course, that the preferences expressed in these percentages are exactly correct, but from the information at hand no evidence can be drawn to support such a belief. In case more exact confirmation is wanted, additional items in the sample would be required to increase the total number of cases, and consequently increase the accuracy of the result.[2]

In the solution as given, one further tacit assumption has been made. This is that there are an unlimited number of brands from which the housewives may make their choice. This assumption eliminates from the methods so far considered many tables of limited choices which often are used in the analysis of markets. The analysis of such limited tables will be found developed in Problem 3.

[1] For proof of formula, see page 21, Corollary.

[2] For discussion of the size of the sample in relation to the accuracy of the result, see page 10ff.

PROBLEM 3. SEVERAL PERCENTAGES: LIMITED CHOICE

Market research frequently is directed at gaining a knowledge of a limited portion of a given market. Thus an investigation in a given city may be limited to consumer preference concerning a particular group of department stores. Naturally this at once may eliminate many specialty stores, mail-order houses, or, in the case of neighboring cities, similar department stores located in such neighboring shopping centers. The housewife being interviewed may not be in the habit of buying at any of the stores named, but very frequently will be found to have a preference for one or another of those named. On the questionnaires the question appears often in the form "At which of the following stores would you prefer to shop?" (There follows a list of the four or five department stores in which the agency is interested and from which the housewife must make her choice.)

The difficulty in situations of this kind arises from the fact that a vote for a particular store or brand or kind of merchandise takes away a vote from some other store or brand. There is thus a negative correlation existing between the groups. This implies at once that the choices for each store or brand are not independent of the choices for the others.

The following case will illustrate the point of attack.

Application — Saxon Advertising Agency[1]

The Saxon Advertising Agency was determining consumer preferences in regard to selected departments of four department stores in a particular city. Conclusions were to be based upon data gathered by interviews with 1,168 housewives. Illustrative of a portion of the data were the votes, classified by stores, showing preferences for the hat department. The figures were as shown in Table 3.

[1] Fictitious name.

Table 3. Preferences for Hat Departments

Store	Number of Votes	Per Cent
Kirkland	671	57.5
Parker	311	26.6
Freeman	138	11.8
Manning	48	4.1

The preference for Kirkland's in comparison with Manning's appeared to be so great that there was little doubt as to the true situation. The percentage difference between Parker's and Freeman's, however, was so much smaller that it was considered essential to find some method of determining whether the observed difference was in fact significant. Because a vote for any one of these stores took away a vote from another, was it not reasonable to suppose that the sampling errors would be materially different from those calculated according to the formulas suggested in Problems 1 and 2? In this event, how could such errors be determined?

Technical Approach

As previously indicated, a vote for any one of the stores presumably takes away a vote from one of the others. The solution of the problem consequently depends upon an assumption of how far the total vote for any particular store might represent the withdrawal of preferential votes from the other stores. The assumption made here is that the total vote withdrawn and given to any one store has been selected from among the votes for the other stores in direct proportion to the positive votes for those stores as shown in the table. It is recognized that in fact this assumption may not be valid. At least, however, it is reasonable and probably it is the best assumption which can be made under the circumstances. The mathematical derivation of the formula obviously will include a consideration of the correlation which exists on the basis of this assumption. As might be expected, a formula which would be of use would give a numerical value for the error applying to the difference between any two percentages in a given limited choice table. It would be expected, therefore, that the formula would include not only the errors to be found in each percentage, but also some expression involving the two percentages.

This formula, expressed in terms of the standard deviation, is[2]

$$\bar{\sigma}_u^2 = \bar{\sigma}_x^2 + \bar{\sigma}_y^2 + \frac{2}{n} p_x p_y.$$

The formula includes the two symbols p_x and p_y which relate to the values of p in the two respective percentages x and y.

[2] For proof of formula, see page 21ff.

Solution

The formula can be applied to test whether the difference in preference for the Kirkland and the Parker stores is significant.

For Kirkland's	*For Parker's*
$p_x = 57.5$	$p_y = 26.6$
$q_x = 42.5$	$q_y = 73.4$
$n = 1168$	$n = 1168$
$\bar{\sigma}_x^2 = \frac{57.5 \times 42.5}{1168} = 2.09$	$\bar{\sigma}_y^2 = \frac{26.6 \times 73.4}{1168} = 1.67$

Hence,

$$\bar{\bar{\sigma}}_u^2 = 2.09 + 1.67 + \frac{2}{1168}(57.5 \times 26.6)$$

$$= 6.38,$$

$$\bar{\bar{\sigma}}_u = 2.53,$$

$$3\bar{\bar{\sigma}}_u = 7.59.$$

The observed difference is

$$57.5 - 26.6 = 30.9.$$

Since this is very much larger than the permissible error, we conclude that the housewives of this city have a very decided preference for Kirkland's hat department. This preference cannot be confused by any chance elements since it expresses a real opinion.

In contrast to this, general observation would seem to indicate that the difference between Freeman's and Manning's, when tested, would prove to be not significant. The test, however, works out as follows:

For Freeman's	*For Manning's*
$p_x = 11.8$	$p_y = 4.1$
$q_x = 88.2$	$q_y = 95.9$
$n = 1168$	$n = 1168$
$\bar{\sigma}_x^2 = \frac{11.8 \times 88.2}{1168}$	$\bar{\sigma}_y^2 = \frac{4.1 \times 95.9}{1168}$
$= .891$	$= .337$

Hence,

$$\bar{\bar{\sigma}}_u^2 = .891 + .337 + \frac{2}{1168}(11.8 \times 4.1)$$

$$= 1.311,$$

$$\bar{\bar{\sigma}}_u = 1.14,$$

$$3\bar{\bar{\sigma}}_u = 3.42.$$

Since the observed difference is 7.7%, it is seen that the percentages obtained by the sample are important. Confidence, therefore, may be placed in the conclusions regarding the relative standing of the stores.

Conclusion

An examination of the numerical work indicates certain interesting conclusions which may be drawn for limited tables:

First, the standard deviation which measures the error in the difference for these limited tables is larger in any case than the corresponding error for unlimited tables.

Second, where the percentages are small and the number of interviews fairly large, as in the case of the comparison of Freeman's and Manning's in the above table, the increase in error chargeable to the limited choice is not very great.

Third, whenever the percentages are relatively large, as in the case of the comparison of Kirkland's and Parker's, even though the sample may be fairly large also, the addition to the standard deviation may be material. If the sample is relatively small, the added portion of the standard deviation chargeable to the effect of the limited choice may be considerable.

Although, with practice, estimates of the error in a single percentage, as in Problem 1, may be made without the accompanying calculations, the addition of other factors, such as the difference between two percentages and the limited choice element, increases the difficulty of making an offhand judgment as to the significance of the results obtained in the sample. For problems of the sort illustrated by the Saxon Advertising Agency it is wiser to rely on calculated values than upon estimates based on judgment only.

PROBLEM 4. AVERAGES FROM SAMPLES

Errors from sampling may appear in connection with averages, as well as in percentage values. Before a discussion of this problem of averages can be undertaken, however, it is necessary to indicate the fundamental difference in the character of the data to be used in this problem and those used in the preceding ones. This discussion turns about the two classes of data known as attributes and variables.

It is generally recognized[1] that data gathered by the sampling process may be of two types. In one the research director may seek to find the presence or absence of a given characteristic or attribute. Thus, as in the preceding problems, he may note the preference of the individual for paste in contrast to some other form of dentifrice, or the preference for a given department store in contrast to other department stores. The data which are accumulated in this way consist simply of an enumeration or count of the number of individuals or homes which possess or do not possess the given attribute. The percentages, therefore, represent merely the ratio of the number of favorable occurrences or the number of unfavorable ones to the total.

A second form which the data may take is an estimate or a measure of some property. Thus, a measurement of the age of an automobile or the age of another piece of mechanical equipment consists not only in noting whether an individual possesses that equipment, but also in measuring some one of its properties. The measurement is not limited necessarily to a measurement of time. It may include a measurement of size or of weight or of some other physical property. The data which are gathered by this process vary from observation to observation. They are known, consequently, as variables. Whenever an average is calculated and that average is based upon data gathered from a sample, the sampling errors of variables must be considered.

Application — Hunt Advertising Agency[2]

In 1934 the Hunt Advertising Agency made a survey of the market for office typewriters. Among other data there was a table which listed for each of the leading manufacturers the age of the typewriters in use in offices. Table 4 presents these data as gathered for Manufacturer A.

Because there was a possibility that the typewriters of any one of the manufacturers included in the survey would be older on the average than those of another manufacturer, it was essential to calculate the average age for each make of machine. Presumably those manufacturers for which the average was greater were losing their relative positions in the market. Since it was well known that typewriters used in offices had a maximum age of not more than 7 or 8 years, the average age as between manufacturers would have shown small numerical differences. Nevertheless, these differences were important because of the large number of typewriters actually in use. It was necessary, therefore, to have some means of estimating the sampling error present in the average as determined from the sample.

Table 4. Manufacturer A

Age of Machine	Number of Machines
New	53
2–3 Years	44
3–5 Years	40
Over 5 Years	42
Total	179

Technical Approach

The technical analysis of a table such as that given in the case involves two distinct problems: (1) the calculation of an average from the data as given; and (2) the estimate of the sampling error present in that average.

The difficulty with the data as given is to be found in the fact that the ages of the machines have been grouped into four classes. Since only four classes are given, any average calculated on the basis of such group data will be open to the criticism of a certain amount of error present. The objection may be raised that in a market investigation the research director will have in hand all the detail covering each machine so that the actual age may be obtained at once for each case. Practically this is not so. In this investigation, for instance, the operators of the machines were asked to state the approxi-

[1] This description follows Yule's definition. See Yule, G. U., *An Introduction to the Theory of Statistics* (London, Charles Griffin and Company, 10th Edition, 1932), p. 7.

[2] Fictitious name.

mate age of the machine. It was doubtful whether an operator could state, except for a very new machine, the exact age of the machine she was using. Moreover, this type of problem occasionally must be solved when only group data are available.

In calculating an average from group data of the type given in the table, the assumption is made that if within each group the middle point be taken as typical of the whole group, the number of cases which are greater than this mid-point value will be the same as the number which are less than the mid-value. Thus, from the table, 44 machines were estimated to be from 2 to 3 years in age. If 2½ years be selected, the assumption is that 22 of the machines will be less than 2½ years old and 22 of the machines will be more than 2½ years old. The result is that within the group the mid-value may be used to replace the actual value for each of the machines. For 2 of the 4 classes given in the table this may be satisfactory. For the other 2 an additional difficulty occurs.

For the first group, listed as new typewriters, it may be assumed that the age varies from 0 to 2 years. Consequently, the mid-age would be 1 year. For the last group of typewriters, those over 5 years old, there enters the very serious difficulty that some typewriters actually in use may be much older than the common life of a machine. Since it is known that few office typewriters are over 8 years old, the mid-point of 6½ years may be accepted.

The data, consequently, would appear as follows:

Adopted Age	Number of Machines	Age × Number = Total Year Machines
1 year.....	53	53
2½ years.....	44	110
4 years.....	40	160
6½ years.....	42	273
Total.....	179	596

The average age is thus $596/179 = 3.33$ years.

The second task is to determine the probable error in this figure as judged on the basis of the errors in sampling. The formulas used in the first three problems are not applicable here because of the nature of the data, which in this case are based upon the measured quantity, age.[1]

If a considerable number of samples similar to that shown in Table 1 were taken from the same market, presumably each average would differ from every other average obtained. The averages of the various samples, if plotted, would tend to group themselves about some central value. It may be assumed that this central point of concentration would approximate fairly closely the true average value. Such a true average value might be defined as the one derived from an exact knowledge of the age of every machine in the whole market. Technically, this may be stated as the value obtained from a "census of the universe". It is not known what this central value is. In addition, the expense of taking a large number of samples makes this course prohibitive. Mathematicians have shown that if the scatter of items about the average in a single sample can be calculated, a measure of the error of that average may be obtained. The formula is[2]

$$\sigma_m = \frac{\sigma_s}{\sqrt{n-1}},$$

where σ_m = the standard deviation of the mean,

σ_s = the standard deviation of the sample of the population for which data are available,

n = the number of items in the sample.

In this formula there are only two quantities for which values need to be obtained. These are σ_s and n. As indicated, the value of σ_s is to be determined from the sample.

The value of σ_s may be calculated from the data given in the table if the assumptions made in calculating the average are accepted. The method of calculating this value of σ_s is derived directly from the definition, which states that the standard deviation is equal to the square root of the average square of the given values corrected by the square of the average of the values. The correction term has the effect of measuring each item from the arithmetic average instead of from some arbitrary origin.

The plan of calculations can be organized so that the amount of computation is reduced to a minimum. For this problem the work may be arranged as follows:[3]

Age	Number of Cases f	df	df'
1	53	53	53
2.5	44	110	275
4	40	160	640
6.5	42	273	1,774.5
	179	596	2,742.5

Hence,

$$\sigma_s^2 = \frac{2742.5}{179} - \left(\frac{596}{179}\right)^2$$

$$= 15.321 - (3.3296)^2$$

$$= 4.235,$$

$$\sigma_s = 2.0579,$$

$$\sigma_m = \frac{2.0579}{\sqrt{n-1}} = \frac{2.0579}{\sqrt{178}}.$$

[1] One exception is noted on page 9.

[2] For proof of formula, see page 23ff.

[3] See Appendix, pages 23–24.

$$= \frac{2.0579}{13.342}$$
$$= .1542,$$
$$3\sigma_m = .45 \text{ years.}$$

Consequently we are likely to have an error of about one-half year in the average age.

From the above data the average age has been found to be 3.33 years. The average age of this make of typewriter as determined from the sample, therefore, presumably will be between 2.87 years and 3.77 years. Practically this means that the average age of the office typewriters of this make probably lies between 2¾ years and 3¾ years.

Conclusion

In the introduction to this problem it was pointed out that a relatively small difference in the average age as between two makes of typewriter might be very significant because of the condition of the large number of individual machines which that average age tended to reflect. It is desirable to carry the inquiry into this phase of the question. If the average age of office typewriters for another manufacturer is 4.4 years with $3\sigma = 0.6$ years, as judged by the sample, how do the average ages of the machines of the two manufacturers compare?

To solve this we use the formula given in Problem 2 which is as applicable in the case of variables as in the case of attributes. The formula is, assuming no correlation present,

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2.$$

Here,

$$\sigma_x^2 = .2378, \qquad \sigma_y^2 = .0344.$$

Hence,

$$\sigma_u^2 = .2722,$$

or

$$\sigma_u = .522,$$
$$3\sigma_u = 1.56 \text{ years.}$$

The observed difference between the ages of the office typewriters of the two manufacturers is $4.4 - 3.3 = 1.1$ years. The error calculated above is $3\sigma = 1.56$ years. The theories as developed indicate that unless this error is exceeded, the observer cannot be certain that another sample might not eliminate the observed difference, which in this case is 1.1 years. In relation to the value of $3\sigma_u$ given above, therefore, this is probably not significant, because the observed difference is less than the calculated error.

This problem is given as typical of those in which an average value has been derived through the sampling method. Such an average value must be the result of measurements which have been taken. The solution gives us an estimate of the limits between which we should expect to find the average value if only we had the data for every typewriter in the field covered by the sample.

It is to be noted that the number of items appears in the denominator of our determination of the error. If the number of items is increased materially, then the size of the error is reduced. This is equivalent to saying that the larger the sample, the more certain is our knowledge about the average age for all office typewriters actually in operation in the district covered by the sample.

In this connection another important observation is that the accuracy increases approximately as the square root of the number of items in the sample. Thus, if it is desired to double the accuracy, which we will assume is equivalent to cutting the possible error in half, we shall have to multiply the number of items in the sample by 4. In order to cut the error to a quarter of the size it will be necessary to multiply the number of items in the sample by 16. This would mean that instead of 179 items we would have to have 2,864 answers to our inquiry. This problem of increased cost in relation to increased accuracy will be discussed later.[1]

Because there is often confusion in selecting formulas for a given problem, it should be emphasized that the decision as to which formula shall be used in determining the error is based first upon the conclusion as to whether the data represent a simple attribute count or a measurement of some physical property. Different formulas covering several different problems for attribute counts have been given in Problems 1 to 3.

[1] See page 11.

SECTION B. SIZE OF SAMPLE

PROBLEM 5

In problems of market research the question often is asked, "How many items must there be in the sample?" The question cannot be answered in the abstract. Information relating to the errors permitted by the conditions of the problem must be known. Thus, in one case, a manufacturer may be entirely satisfied to secure results which are accurate to within 10% either way. In other cases, such as setting of insurance premiums, it is desirable to be able to estimate the pure premium with the maximum of accuracy.

The question "How many items are necessary in a sample?" consequently must be deferred until the question "How accurate is it necessary to know the result?" is answered. Once the accuracy needed in a given problem has been determined, it is perfectly possible on the basis of the formulas set up for the problems of Section A to give the number of items in a random sample which will be necessary.

Application — Douglas Company[1]

The Douglas Company, which manufactured playing cards, was considering the suitability of selling by mail. Certain mailing lists were available, including lists of club members and college graduates, as well as commercial lists. One of the officers of the company ventured a guess that if 20% of the names represented bridge players, it would be profitable to circularize the lists. It was estimated that possibly 30% of the persons in the available lists were bridge players.

To assist them in making a decision the executives undertook to check these estimates by means of a sample questionnaire. The problem then became one of determining the number of questionnaires necessary.

Technical Approach

No new technique is needed, since the solution depends upon the formula developed in Problem 1.

Solution

The first step in solving the problem is to decide from the information given what the permissible error is. In this particular case the assumed value of p is 30%, since this appeared to be the most probable suggestion. Apparently the Douglas Company will be satisfied even if 20% of the individuals included in the lists in hand represent bridge players. This is 10% less than the 30%. Consequently, the permissible error is 30% less 20% or 10%. Hence $3\bar{\sigma}=10\%$. With reference to the technical approach involved in the solution to Problem 1, the formula given was

$$\bar{\sigma}=\sqrt{\frac{pq}{n}}.$$

A range equal to three standard deviations was used. Consequently, here the following values hold:

$$\bar{\sigma}=3.33\%,$$
$$p=30\%,$$
$$q=70\%.$$

Substituting in the formula, it will be found that $n=189$.

Suppose replies are obtained from 200 persons. If 62 of these are bridge players, the value of p is 31%. This approximately agrees with the preliminary estimate. Since by assumption $3\bar{\sigma}=10\%$, the manufacturer can be practically certain that at least 20% are bridge players.

The original estimate, however, may be quite erroneous. Suppose, for example, that only 40 of the replies are from bridge players. This makes $p=20\%$. Since presumably the manufacturer had good reasons for believing originally that $p=30\%$ we have an apparent contradiction which must be cleared up. The original estimate may be wrong or the sample taken may be of such an unusual character that the conclusions derived from it are wrong. The first of these assumptions seems the more reasonable to make since three standard deviations were used in determining the number of items to be taken in the sample. On the other hand, unusual events sometimes do happen.

If it is assumed that the original assumption of 30% for p is in error and that the value of 20% originally should have been taken, in place of 189 items, 144 might have been satisfactory. This may be shown as follows:

[1] Fictitious name.

$$\bar{\sigma} = \sqrt{\frac{pq}{n}}$$

$\bar{\sigma} = 3.33\%$,
$p = 20\%$,
$q = 80\%$.

Substituting in the formula, it will be found that $n = 144$.

A second element was present in the problem, however, which needs further consideration. The manufacturer estimated that he would need at least 20% in order to make his proposition commercially satisfactory. It has been assumed, however, that the margin was 10% either way. To overcome this difficulty, two possibilities are available. Both of these are based upon the necessity of determining the value of p more accurately.

One possible approach is to be found in the assumption of 20% as the accurate value of p, with the limitation that the error shall not be more than 2% either way. A calculation similar to that above shows that in order to determine the value of p with this degree of accuracy the sample would have to be increased to 3,600 returns actually received. If it were desired to determine the value of p to within 1%, the sample would have to be increased to 14,400. This might be a prohibitive number.

A second possible approach is to assume that the value of p determined from the small sample is somewhat unusual and that the true value would be nearer 25%. A calculation shows that an allowance of 5% either way would require 675 replies in the sample. Therefore, increasing the size of the sample somewhat will lead to a decision as to the next step in this uncertain case.

The solution of the problem obviously turns about a process of trial and error. The initial step is to assume the value of p which is believed to be most reasonable. The results obtained from the sample will determine whether the guess is entirely satisfactory. If the value of p obtained from the sample approximates very closely a limiting value which would make a given project commercially undesirable, as in the case of the Douglas Company, then a further trial must be made. This must be repeated so long as there is doubt in regard to the result obtained.

In comparison with the procedure outlined above, the practical rule used by many market investigators in connection with sampling is at times unnecessarily cumbersome. This rule involves successive samples of 100 or 200 items and the accumulation of the results until the value of p appears to have become stabilized. It should be noted that this rule leaves open entirely the question of the degree of stabilization which is necessary for the problem under consideration.

The procedure outlined for the Douglas Company makes it possible to estimate in advance the cost of a particular investigation. This assumes that an approximate idea of the value of p is available from other market studies.

There is a further advantage in pre-planning. As soon as the accuracy desired in the results has been determined by the conditions of the problem and the cost estimated, the business executive should be able to decide whether the results are worth the price. Not infrequently have extensive market investigations been undertaken at a very substantial cost when the information sought, in fact, could be obtained from a relatively small sample costing at most a very few thousand dollars. Conversely, investigations have been undertaken in which the final figures were expected to be determined so accurately that costs would run into hundreds of thousands of dollars in order to obtain a result which had a commercial value of only a fraction of that price.

Table for Size of Sample to Have a Given Reliability

In planning market researches a number of estimates of the type indicated above frequently will have to be made. For such purposes a table of values will be found convenient for reference. This is given on pages 12 and 13.

Certain peculiar characteristics of the table will be recognized. The two numbers at the top of each column obviously total 100%; one value is for p and the other is for q, at the option of the user. Moreover, it will be noticed that the values on the top row go only to 50%. For values greater than 50% the lower row should be used. The remaining of the two values, after p has been selected, will be the value of q. The reason that this can be done is that the formula is symmetrical for p and q.

The column on the left-hand side of the table states the error in per cent. It will be noted that these values are not all equally spaced from one another, the lower values being given at smaller intervals than the larger ones. The figures in the body of the table state the number of items in each case corresponding to the values of p and q in the error selected. Two illustrative examples will indicate the use of the table.

A market research agency desires to take a sample for which it estimates the value of p will be approximately 25%. It wishes to determine this within limits of 1% either way. How many items will be required in the sample? Answer: Under the column headed $\frac{25}{75}$ and on the line opposite 1.0 in the column headed

Size of Sample Necessary to be Practically Sure (i.e. basis of $\pm 3\bar{\sigma}$) of Accuracy Within Given Limits

Formula: $n = \frac{9\,pq}{(3\bar{\sigma})^2}$

Limits $\pm 3\bar{\sigma}$ (in %)	Values of p and q^* (in %)						
	1 99	2 98	3 97	4 96	5 95	10 90	15 85
.1	89,100	176,400	261,900	345,600	427,500	810,000	1,147,500
.2	22,275	44,100	65,475	86,400	106,875	202,500	286,875
.3	9,900	19,600	29,100	38,400	47,500	90,000	127,500
.4	5,569	11,025	16,369	21,600	26,719	50,625	71,719
.5	3,564	7,056	10,476	13,824	17,100	32,400	45,900
.6	2,475	4,900	7,275	9,600	11,875	22,500	31,875
.7	1,818	3,600	5,345	7,053	8,724	16,531	23,418
.8	1,392	2,756	4,092	5,400	6,680	12,656	17,930
.9	1,100	2,178	3,233	4,267	5,278	10,000	14,167
1.0	891	1,764	2,619	3,456	4,275	8,100	11,475
1.5	396	784	1,164	1,536	1,900	3,600	5,100
2.0	223	441	655	864	1,069	2,025	2,869
2.5	143	282	419	553	684	1,296	1,836
3.0	99	196	291	384	475	900	1,275
3.5	73	144	214	282	349	661	937
4.0	56	110	164	216	267	506	717
4.5	44	87	129	171	211	400	567
5.0	36	71	105	138	171	324	459
6.0	25	49	73	96	119	225	319
7.0	18	36	53	71	87	165	234
8.0	14	28	41	54	67	127	179
9.0	11	22	32	43	53	100	142
10.0	9	18	26	35	43	81	115
15.0	4	8	12	15	19	36	51
20.0	2	4	7	9	11	20	29
25.0	1	3	4	6	7	13	18
30.0	1	2	3	4	5	9	13
35.0	.7	1	2	3	3	7	9
40.0	.6	1	2	2	3	5	7

* If either number in the column heading is selected for p, the other is equal to q, because the sum of p and q equals 100%.

"Limits" there is found the value 16,875. This is the number of items in the sample.

The director of research believes that the proportion of families constituting the market for a product which he is interested in promoting consists of not more than 10% of the total number. He will be entirely satisfied if the results are accurate to within 3%. How many items must be in the random sample? Answer: 900.

Size of Sample Necessary to be Practically Sure (i.e. basis of $\pm 3\bar{\sigma}$) of Accuracy Within Given Limits (continued)

Formula: $n = \frac{9pq}{(3\bar{\sigma})^2}$

Values of p and q* (in %)							Limits $\pm 3\bar{\sigma}$ (in %)
20 80	25 75	30 70	35 65	40 60	45 55	50 50	
1,440,000	1,687,500	1,890,000	2,047,500	2,160,000	2,227,500	2,250,000	.1
360,000	421,875	472,500	511,875	540,000	556,875	562,500	.2
160,000	187,500	210,000	227,500	240,000	247,500	250,000	.3
90,000	105,469	118,125	127,969	135,000	139,219	140,625	.4
57,600	67,500	75,600	81,900	86,400	89,100	90,000	.5
40,000	46,875	52,500	56,875	60,000	61,875	62,500	.6
29,388	34,439	38,571	41,786	44,082	45,459	45,918	.7
22,500	26,367	29,531	31,992	33,750	34,805	35,156	.8
17,778	20,833	23,333	25,278	26,667	27,500	27,778	.9
14,400	16,875	18,900	20,475	21,600	22,275	22,500	1.0
6,400	7,500	8,400	9,100	9,600	9,900	10,000	1.5
3,600	4,219	4,725	5,119	5,400	5,569	5,625	2.0
2,304	2,700	3,024	3,276	3,456	3,564	3,600	2.5
1,600	1,875	2,100	2,275	2,400	2,475	2,500	3.0
1,176	1,378	1,543	1,671	1,763	1,818	1,837	3.5
900	1,055	1,181	1,280	1,350	1,392	1,406	4.0
711	833	933	1,011	1,067	1,100	1,111	4.5
576	675	756	819	864	891	900	5.0
400	469	525	569	600	619	625	6.0
294	344	386	418	441	455	459	7.0
225	264	295	320	338	348	352	8.0
178	208	233	253	267	275	278	9.0
144	169	189	205	216	223	225	10.0
64	75	84	91	96	99	100	15.0
36	42	47	51	54	56	56	20.0
23	27	30	33	35	36	36	25.0
16	19	21	23	24	25	25	30.0
12	14	15	17	18	18	18	35.0
9	11	12	13	14	14	14	40.0

* If either number in the column heading is selected for p, the other is equal to q, because the sum of p and q equals 100%.

SECTION C. NOTES ON SAMPLING PROBLEMS

Approaching the Problem of Market Research

In approaching market research undertakings the function of the field work in relation to the data and the organization of questionnaires should be kept clearly in mind. The statement often is made that field work is undertaken in order to secure information. Such a statement is equivalent to saying that an inventor carries on laboratory work in order to invent something, the nature of which is not altogether clear.

Essentially market research should be planned to verify or disprove assumptions that have been developed out of the experience of the manufacturer or advertising agency, or to seek the answers to questions which the manufacturer has suggested. In turn, these assumptions and questions must be formulated in such a manner that the field investigation verifies or disproves them. This may seem to be obvious. Yet from the results as submitted in many investigations it is apparent that the implications in these statements have been realized only partially.

So far as the statistical data are concerned, schedules and classifications commonly are set up. This work is done presumably on the basis of experience and/or advice of the market consultant. The data gathered in the investigation should be analyzed for the purpose of deciding whether the assumptions made in the various classifications hold in fact. It is toward this end that the problems as outlined assist in reaching decisions about the assumptions which have been made. Some illustrations will make clear the problem.

In certain research, for example, a market agency assumed that the use of a specified household utensil would vary as between economic classes of the population. These were

Income Class A — very wealthy,
B — upper middle class,
C — lower middle class,
D — foreign born population.

The data as obtained in a sample showed that there was no significant difference between classes A and B, even though the sample gave an observed difference of some 5% or 6%. As between classes B and C the difference observed in the sample was so large that apparently there was a significant distinction to be made between families falling in either of these two classes. Finally, the number of items in the sample taken for Class D was so small, with the consequent errors of sampling so large, that no significant statement could be made in regard to this class of the population.

Unfortunately, the whole table was submitted to the client. In spite of the fact that different percentages were observed in the sample for each class respectively of the population, the data really indicated only one fact. This was that apparently there was a difference in consumer demand between Classes B and C of the population. The conclusion as reported to the client should have stated that for his purposes the population could have been divided into two groups, namely, the upper middle class and wealthy, and the lower middle class and foreign born.

In Problem 2, which shows preferences for brands, the same type of question occurs. So far as the statistical data presented are concerned, there is no evidence to support the conclusion that there are different degrees of consumer demand for the different brands except as between widely separated brands. For the purposes of a table such as that given in Problem 2, the market research indicates that the assumption of a different consumer demand for the different brands does not hold for the first two or three brands. Apparently these may be grouped to indicate a single type of demand. Those which follow fall roughly into a second group, while those which indicate a smaller consumer preference comprise a poor third. It is recognized in a problem such as this, where the percentages are grouped so closely, that it is practically impossible to draw sharp lines of division. If this is the hypothesis to be tested, then a materially increased number of items in the sample should be taken.

Random Sample

The solution of the problems presented in this study has been based upon the theory that the samples were selected in a random manner. This implies that absolutely no personal interest or bias entered into the selection of individuals from whom answers were to be obtained.

In application, much depends upon the degree of homogeneity or the uniformity of the market which is

being examined. Every individual differs from every other individual in tastes and in desires, so that in the strictest sense of the word there is no uniformity in the population which is being examined. From a practical point of view this is not the important question, since in their attitude toward particular goods whole communities may possess fairly uniform characteristics. Thus, if the research is directed at testing the market for an electrical household appliance, obviously the population to be examined will not include those homes which are not wired for electricity. Or, again, if the device be a heater for an automobile, presumably the population in the extreme southern parts of the United States or in the southern Pacific Coast region will not be included in the examination.

The problem becomes very much more complex when it is desired to test an urban market population. Here differences in economic condition and differences in social habits and customs, reflecting differences in income, may or may not cause the elimination of certain sections of the population or certain districts of the city examined.

The conclusion is obvious that the market analyst must decide specifically what is his objective and what is the population, or the portion of the population, from which he wishes to secure evidence.

The adjective representative has been used with reference to a sample. This word tends to introduce confusion in the thinking concerning the method of selecting a sample. Among other research papers on the subject a recent article by J. Neyman [1] indicates that the method of selecting a representative sample may be divided into two broad groups, namely, the method of random sample and the method of purposive selection. The problem here will be dismissed with a simple statement that apparently random samples taken from stratified groups of the population will yield sampling data to which the formulas developed will apply.

Breaking Down a Sample

Closely related to the question of the number of items which are to be included in the sample is the question of the breakdown that ultimately is to be made. Sufficient allowance in the size of the original sample must be made so that there will be a large enough number of items in the smallest breakdown. Otherwise the small sub-classes will give results so inaccurate that their contribution to the whole study will be negligible. The emphasis must be placed upon distinguishing between a sub-class which is independent and one which comprises a portion of a larger group. An example of a sub-class which is independent would be the group of families within a certain income class. This sub-class is independent of other income groups of families. An example of a sub-class not independent would be the number of families within a certain income group who possessed a radio as compared with those who owned an automobile.

Frequently confusion arises in the mind of the investigator as to the value of n, the number of items which shall be selected in a given problem. No rule can be suggested. There is, however, a method of approach which will help the investigator to make a proper decision.

The value selected for n should reflect the total representative population from which the random sample is drawn. If the random sample is to be broken down into sub-classes, then the total number of interviews in each sub-class should be the value for n. In such cases it is assumed that each sub-class has been selected in a random manner.

On the other hand, if votes are taken for several objects, then the total n should be for the whole group. Thus, in Problem 2 above, there were a number of brands of toilet soap, any one of which might have been selected. The total number of replies was 534. The random chance that any one brand would be selected is based upon the 534 replies and not on the number of replies for any one particular brand.

Methods of Selecting Samples

If it be granted that a sample must be a random one, the question may be asked "How can such a sample be selected from the whole population?"

A first answer might be that the investigator should select the homes at random on some street or streets of a given city or town. Such a procedure, however, is open to unconscious bias. Attractive sections are very likely to receive more attention than the unattractive ones. Along a particular street certain houses may be so designed that they appeal to the particular investigator in charge of making this haphazard or random selection. Finally, those homes which might appeal unconsciously to one, very likely would have no appeal to another investigator.

A method which often is suggested is that the names of the families be listed on slips of paper which subsequently will be shuffled and drawn at random. Practically this procedure is open to two objections. The first is the expense and time involved in performing the task. The second is the difficulty that slips of paper have a tendency to stick together, so that the shuffling is not a wholly random one. This last difficulty might be avoided by enclosing each name in a gelatin capsule

[1] "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 1934, pp. 558–625.

in order that out of a drum containing capsules for the total population the sample might be drawn. This, however, does not remove the objection of expense in performing the task.

Another procedure is to take the directory or telephone book as including the list of names of all residents. Every fifth, tenth, twentieth, or hundredth name then is selected. It is to be noted that such a sample is biased in the direction of the weighting of names, which often reflect racial characteristics if the sample is drawn from a city largely populated by the foreign born or their immediate descendants. It is also to be remembered that in selecting names from a telephone directory, the tendency is to cover that portion of the market which has a higher buying power. For some goods, the use of the telephone directory will constitute a limitation of the population which it is desired to sample.

Still another procedure is to number the list of residents and then to select, at random, figures from a table, such as a table of logarithms, using the last three or four digits as the case may be. Presumably this would be fairly satisfactory, but it involves the expense of numbering the lists in the directory or telephone book.

A table of completely random numbers has been compiled by L. H. C. Tippett and published by the Cambridge University Press, London. This gives a list of 10,000 consecutive numbers arranged in a completely random order.

From the foregoing it will be judged that the selection of a random sample is not easy. The limitations of funds, as well as the practical difficulties of selection, make compromises necessary. If the investigator is expecting unconscious bias to enter at every point, he will at least make an attempt to select a sample which will be as nearly random in character as his funds or time or opportunity permit.

The Standard Deviation

The standard deviation in the problems developed here is used as a basis for the measure of the error. An understanding of its significance is important.

Consider a measured quantity, such as the time it takes a business man to go from his home to his office. From a number of observations which are equivalent to his experience he will tell you about how much it takes in time. This figure represents the average, which he will agree at once is not within a fraction of a second of the same time every morning. On some days the journey will take a little more time and on others it will take less time than the average. This variation in time from the average is significant.

If the variation in time in our illustration is rather large, we say that the man's performance tends to be erratic. On the other hand, if the time from trial to trial varies but slightly, we say that his performance is highly consistent. It is relatively easy to set up some kind of measure which will indicate how many times a given departure from the average actually has occurred. It might be said in fact that two-thirds of the time the business man has come within such and such limits or has not varied more than so many minutes from his average time.

The measure of consistency suggested in the foregoing paragraph technically is known as the standard deviation. In certain kinds of distributions it includes two-thirds of the trials on either side of the average, exactly as in the case of the business man measuring the time consumed in getting from his home to his office. The method of calculating the standard deviation will be found in the Appendix, page 23. References to any of the standard texts on statistics will give further information about this measure. The important fact is that the standard deviation is a measure of consistency of performance, and, as developed for use in the problems of market research, is used as a measure of the error.

It will be noted that a range of three standard deviations is used instead of one standard deviation. Experience shows that for practical purposes it apparently is safe to say that three standard deviations will be exceeded only once in 100 trials. This does not mean that some unusual combination of circumstances may not give a deviation much larger than three standard deviations from the average, but the relative probability is small. That the work must be based on experience and not upon mathematical logic is explained in the following note on *a priori - a posteriori* probability. In some cases two standard deviations apparently are ample to meet the circumstances of the problem. In general, however, it is believed wise to use three rather than two.

A Priori – A Posteriori Probabilities

Since the whole theory of sampling is based upon the theory of probability, it is pertinent to review briefly the philosophy which lies behind the method of attack.

In games of chance, such as those using dice, the theoretical chance that a given combination will turn up may be calculated in advance by a mathematical process. This calculation is based essentially upon the mechanical features of the articles used in the play. Thus, a die has only six faces. Consequently, the theoretical chance of obtaining an ace is 1 in 6. It is assumed in making such a statement that all faces have an equal chance of turning uppermost. This involves the

assumption that the die is accurately cut and is of homogeneous material, i.e. not loaded, and also the assumption that it is handled uniformly in play.

Whenever two dice are involved the probability becomes a compound one. Thus, the probabilities of each are calculated separately and are then combined. Whether the combination is made by a process of multiplication, or by a process of addition to give the compound event, will depend upon the character of the problem. A similar approach holds true for the distribution of cards in a game of bridge. Thus it has been calculated that the chance of securing a perfect hand of 13 cards of a set from a completely random shuffle is 1 in 158,753,389,900.

Because the probabilities can be calculated in advance for games of chance such as dice or the deal in a game of bridge, the probabilities relating to such games are known as *a priori* probabilities. Obviously, games of chance of this sort constitute a distinct class.

In problems relating to the actions of human beings no such preliminary calculations can be made. Among problems of this sort is the chance of a person's living a year or any given number of years, or the problems of casualty insurance, or the chances of a given condition in a market being obtained from the use of a sample. For all the problems of this kind, where the probabilities cannot be calculated in advance, the attempt in practice is to estimate the chance of success or failure from a collection of experience. Such experience is reduced to figures and the data are interpreted with the hope and with the expectation that they would approximate the true values if they were but known.

Probabilities of this kind are known as *a posteriori* probabilities since they are chances whose values are estimated only on the basis of fragmentary experience. Nevertheless mathematicians and statisticians have been able to work out approaches to this last group of problems, and, in spite of the logical shortcomings, have demonstrated that with the use of these approaches practical affairs of mankind can be handled.

In building up the mathematical approach to the second group of problems it often is found helpful to proceed by studying simpler problems of the first kind (*a priori*) which have been indicated. In the development of the mathematical formulas which are appended, the reader will note that at various points such assumptions have been made.

APPENDIX

MATHEMATICAL NOTES

The following notes develop the mathematical formulas used in solving the problems. There is a twofold reason for including them. The first is that they complete the material already presented, by giving mathematical proof for the theorems used; the second is that they may provide the necessary groundwork for those readers who take pleasure in the mathematical approach to problems.[1]

The problems themselves form an increasingly complicated set of cases. The notes added here follow the same order.

PROBLEM 1. DERIVATION OF FORMULAS FOR THE ARITHMETIC MEAN AND STANDARD DERIVATION

If p represents the chance of success and q the chance of failure of an event, then we have by definition

$$p+q=\text{certainty}$$
$$=1.$$

If there are n independent events, the chance that exactly r of them out of the total n will happen is

$${}_nC_r p^r q^{n-r},$$

where ${}_nC_r$ is the customary symbol for number of combinations of n things taken r at a time.

It is well known that this expression is the rth term of the expansion of the binomial $(p+q)^n$. If the expansion is made beginning with q^n,

$$(q+p)^n=q^n+nq^{n-1}p+\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2+\cdots+p^n.$$

The successive terms of this expansion will give the choices of 0, 1, $2\cdots n$ successes. These algebraic data may now be arranged in tabular form for the computation of the arithmetic mean and the standard deviation. The organization of the work follows the usual procedure.

Frequency f	Number of Successes d'	fd'	$f(d')^2$
q^n	0	0	0
$nq^{n-1}p$	1	$nq^{n-1}p$	$nq^{n-1}p$
$\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2$	2	$n(n-1)q^{n-2}p^2$	$2n(n-1)q^{n-2}p^2$
$\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}q^{n-3}p^3$	3	$\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$	$\frac{3n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
p^n	n	np^n	n^2p^n
$\Sigma f=N=1$		$\Sigma fd'=np$	$\Sigma f(d')^2=np[1+p(n-1)]$

[1] See Bowley, A. L., *Elements of Statistics* (New York, Scribner's, 1920).
Jones, D. C., *First Course in Statistics* (London, Bell & Sons, 1921).
Yule, G. U., *Theory of Statistics* (London, Griffin and Company, 1932).

The derivation of the sums is as follows: Obviously the value of $\Sigma f = 1$, since the terms in this column are the terms in the expansion of $(q+p)^n$.

But $$q+p=1.$$

Hence, $$(q+p)^n=1=N=\Sigma f.$$

To find the value of $\Sigma fd'$ we factor from each term the quantity np. This leaves the series

$$q^{n-1},$$
$$(n-1)q^{n-2}p,$$
$$\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2,$$
$$\cdots$$
$$\cdots$$
$$p^{n-1}.$$

The sum of these terms is

$$(q+p)^{n-1}.$$

But this has the value unity, since

$$q+p=1.$$

Hence,

$$\Sigma fd'=np,$$

since np was factored from each term.

The sum of the terms in the final column is obtained by first factoring the value np. This gives

$$\Sigma f(d')^2=np\left[q^{n-1}+2(n-1)q^{n-2}p+\frac{3(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots+np^{n-1}\right].$$

The terms within the brackets may now be divided into two parts, giving

$$\Sigma f(d')^2=np\left[\left(q^{n-1}+(n-1)q^{n-2}p+\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots+p^{n-1}\right)\right.$$
$$\left.+\left((n-1)q^{n-2}p+\frac{2(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots+(n-1)p^{n-1}\right)\right].$$

The first series in parentheses is the expansion of

$$(q+p)^{n-1}=1.$$

If $(n-1)p$ be factored from the second parentheses the remaining terms are the expansion of

$$(q+p)^{n-2}=1.$$

Hence we have

$$\Sigma f(d')^2=np[1+(n-1)p].$$

Our arithmetic mean by definition is

$$\text{Arithmetic Mean}=M_a=\frac{\Sigma fd'}{\Sigma f}$$

$$=\frac{np}{1}=np.$$

The standard deviation is

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\Sigma f(d')^2}{\Sigma f} - \left(\frac{\Sigma f d'}{\Sigma f}\right)^2}$$
$$= \sqrt{np[1 + p(n-1)] - n^2p^2}$$
$$= \sqrt{np(1-p)}$$
$$= \sqrt{npq}.$$

The relationship between σ and $\bar{\sigma}$ is as follows. Since $\bar{\sigma}$ is defined as expressing σ as a percentage of the number of observations n,

$$\frac{\sigma}{n} = \bar{\sigma}.$$

But since

$$\sigma = \sqrt{npq},$$
$$\frac{\sigma}{n} = \frac{\sqrt{npq}}{n} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}.$$

Hence,
$$\bar{\sigma} = \sqrt{\frac{pq}{n}}.$$

PROBLEM 2. DERIVATION OF FORMULA FOR SUM OR DIFFERENCE OF TWO VARIABLES

We desire here to prove the formula

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2 \pm 2r\sigma_x\sigma_y.$$

Let x = the first variable,

y = the second variable,

and $u = x \pm y$.

Then
$$u^2 = x^2 \pm 2xy + y^2.$$

Hence,
$$\Sigma u^2 = \Sigma x^2 \pm 2\Sigma xy + \Sigma y^2,$$

and
$$\frac{\Sigma u^2}{n} = \frac{\Sigma x^2}{n} \pm \frac{2}{n}\Sigma xy + \frac{\Sigma y^2}{n}. \qquad (1)$$

By definition the standard deviation is to be measured from the arithmetic mean. The above equation must be corrected if x, y and u are measured from an origin other than that of the mean.

Since
$$u = x \pm y,$$
$$\frac{\Sigma u}{n} = \frac{\Sigma x}{n} \pm \frac{\Sigma y}{n}.$$

Hence,
$$\left(\frac{\Sigma u}{n}\right)^2 = \left(\frac{\Sigma x}{n}\right)^2 \pm \frac{2}{n^2}\Sigma x \Sigma y + \left(\frac{\Sigma y}{n}\right)^2. \qquad (2)$$

Subtracting (2) from (1) we have

$$\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 \pm 2\left[\frac{\Sigma xy}{n} - \frac{\Sigma x \Sigma y}{n^2}\right] + \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2.$$

Whence,

$$\sigma_u^2=\sigma_x^2\pm 2\left[\frac{\Sigma xy}{n}-\frac{\Sigma x\Sigma y}{n^2}\right]+\sigma_y^2.$$

Now since the coefficient of correlation

$$r=\frac{\Sigma xy-\frac{\Sigma x\Sigma y}{n}}{n\sigma_x\sigma_y},$$

we have

$$\sigma_u^2=\sigma_x^2\pm 2r\sigma_x\sigma_y+\sigma_y^2.$$

Corollary:

If x and y are independent

$$r=0.$$

Hence,

$$\sigma_u^2=\sigma_x^2+\sigma_y^2.$$

Similarly, for percentages,

$$\bar{\sigma}_u^2=\bar{\sigma}_x^2+\bar{\sigma}_y^2.$$

PROBLEM 3. DERIVATION OF FORMULA FOR LIMITED NUMBER OF ATTRIBUTES

Derivation of the formula

$$\bar{\bar{\sigma}}_u^2=\bar{\bar{\sigma}}_x^2+\bar{\bar{\sigma}}_y^2+\frac{2}{n}p_xp_y.$$

It has been shown above that if

$$u=x-y,$$

$$\sigma_u^2=\sigma_x^2+\sigma_y^2-2r\sigma_x\sigma_y.$$

For tables of limited choices it is necessary to find the value of r between the variation in two frequencies of a sample distribution. To derive this value let the distribution have in one frequency f_s a variation of x votes corresponding to a variation of y votes in another frequency f_t. Then if the total frequency is n, $n-f_s$ votes are included in classes other than the f_s class. Hence if x votes are to be added to the f_s class and if it is assumed that these are taken from the other classes in proportion to their frequencies, we have for the contribution from the f_t class the proportion $\frac{f_t}{n-f_s}$ and the number $\frac{f_t}{n-f_s}x$. Therefore,

$$y=-\frac{f_t}{n-f_s}x. \tag{1}$$

Now the value of r sought is

$$r_{xy}=\frac{\Sigma xy}{N\sigma_x\sigma_y} \tag{2}$$

where N is the total number of samples and where it is assumed that the variates x and y are measured from their means f_s and f_t respectively. Also, since for the f_s class in the sample distribution,

$$\sigma_x^2=npq,$$

and

$$p=\frac{f_s}{n},$$

$$q=1-\frac{f_s}{n},$$

we have

$$\sigma_x^2=f_s\left(1-\frac{f_s}{n}\right)$$

$$=\frac{nf_s-f_s^2}{n}.$$

From (1)

$$xy=-\frac{x^2f_i}{n-f_s}$$

$$=-\frac{1}{n}\frac{x^2f_if_s}{\frac{nf_s-f_s^2}{n}}$$

$$=-\frac{1}{n}\frac{x^2f_if_s}{\sigma_x^2}.$$

Hence,

$$\Sigma xy=-\frac{1}{n}\frac{f_if_s\Sigma x^2}{\sigma_x^2}.$$

But

$$\sigma_x^2=\frac{\Sigma x^2}{N}.$$

Hence,

$$\Sigma xy=\frac{1}{n}\frac{f_if_sN\sigma_x^2}{\sigma_x^2}$$

$$=-\frac{N}{n}f_if_s.$$

Substituting in (2),

$$r_{xy}=-\frac{1}{n}\frac{f_if_s}{\sigma_x\sigma_y}.$$

Substituting this value in

$$\sigma_u^2=\sigma_x^2+\sigma_y^2-2r\sigma_x\sigma_y,$$

we have

$$\sigma_u^2=\sigma_x^2+\sigma_y^2+\frac{2}{n}f_if_s.$$

When we use percentages of the total frequency n, call

$$\frac{f_i}{n}=p_x,$$

$$\frac{f_s}{n}=p_y,$$

so that dividing the above equation by n^2, we have

$$\bar{\sigma}_u^2=\bar{\sigma}_x^2+\bar{\sigma}_y^2+\frac{2}{n}p_xp_y.$$

PROBLEM 4. CALCULATION OF THE STANDARD DEVIATION FROM A FREQUENCY DISTRIBUTION

If d = the deviation of any item from the mean M,
c = the deviation from the mean M of an arbitrary origin selected for convenience in calculation,
d' = the deviation of any item from the arbitrary origin c,
n = the number of items,
σ = the standard deviation,
then by definition

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}}.$$

Also from the definitions

$$d' = d - c.$$

Whence,

$$d'^2 = d^2 - 2cd + c^2,$$

$$\Sigma d'^2 = \Sigma d^2 - 2c\Sigma d + nc^2.$$

Since d is measured from the mean M,

$$\Sigma d = 0.$$

Hence,

$$\Sigma d'^2 = \Sigma d^2 + nc^2,$$

or

$$\Sigma d^2 = \Sigma d'^2 - nc^2.$$

Therefore,

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{\Sigma d'^2}{n} - c^2}.$$

When we calculate from a frequency distribution with the frequency in any class equal to f, the formula becomes

$$\sigma = \sqrt{\frac{\Sigma f d'^2}{n} - c^2}.$$

If for brevity we set

$$\frac{\Sigma f d'^2}{n} = S^2,$$

$$\sigma = \sqrt{S^2 - c^2}.$$

Calculation by Short-Cut Method
Standard Deviation of Age of Automobiles*

Manufacturer A

Age Group Years Class Interval 2 Years	Midpoint m	Frequency f	Deviations d'	fd'	$f(d')^2$
0– 1.9	1	94	−2	−188	376
2– 3.9	3	247	−1	−247	247
4– 5.9	5	356	0	0	0
6– 7.9	7	268	+1	+268	268
8– 9.9	9	195	+2	+390	780
10–11.9	11	54	+3	+162	486
12–13.9	13	10	+4	+ 40	160
		$N = 1224$		+860	$\Sigma f(d')^2 = 2317$
				−435	
				$\Sigma fd' = +425$	

*Data fictitious to illustrate process only.

Hence,

$$S^2=\frac{2317}{1224}=1.893,$$

$$c=\frac{425}{1224}=0.347,$$

$$c^2=0.120,$$

$$\sigma^2=1.893-0.120,$$

$$=1.773,$$

$$\sigma=1.33.$$

PROBLEM 5. DERIVATION OF FORMULA FOR ERROR OF THE MEAN

It was proved above (page 21) that if x and y are independent,

$$\sigma_u^2=\sigma_x^2+\sigma_y^2.$$

In a similar manner it might be proved for

$$u=x+y+z+\cdots+w,$$

that

$$\sigma_u^2=\sigma_x^2+\sigma_y^2+\sigma_z^2+\cdots+\sigma_w^2.$$

If now $x, y, z \cdots w$ represent n measurements, all of which are subject to the same law of error,

$$\sigma_x=\sigma_y=\sigma_z=\cdots=\sigma_w.$$

Let any one have the value σ,
that is,

$$\sigma_x=\sigma_y=\sigma_z=\cdots=\sigma_w=\sigma.$$

Then if the n measurements are independent,

$$\sigma_u^2=n\sigma^2,$$

or

$$\sigma_u=\sigma\sqrt{n}.$$

Now if instead of u expressed as the sum of measurements we took the average, then u as well as $x, y, z \cdots w$ would each be divided by n. Then σ_u would be divided by n and in fact would be the standard deviation of the mean $=\sigma_M$.
Hence,

$$\frac{\sigma_u}{n}=\sigma_M=\frac{\sigma\sqrt{n}}{n}=\frac{\sigma}{\sqrt{n}}.$$

If the sample is large the σ of the sample may be used as the standard deviation of the universe. When, however, the sample is small, researches[1] have shown that a better value for the denominator is $\sqrt{n-1}$.

As will be noted, the assumption has been made that the measurements are all subject to the same law of error. It is recognized that this does not always hold. Nevertheless, experience seems to show that the formula holds reasonably well for all but extreme cases. For researches covering cases where the distribution is not normal, the reader is referred to the statistical journals.

[1] See, for example, the paper by "Student," *Biometrica*, 1908, pp. 1–25.

BUREAU OF BUSINESS RESEARCH: BULLETINS IN PRINT — *Continued*

DRUG — WHOLESALE

No. 50. Operating Expenses in the Wholesale Drug Business in 1924 ..... 50 cents
No. 46. Operating Expenses in the Wholesale Drug Business in 1923 ..... 50 cents

DRY GOODS — WHOLESALE (Southern)

No. 45. Operating Expenses in the Wholesale Dry Goods Business in the South in 1923 ..... 50 cents

GROCERY — RETAIL (*See also* CHAIN STORES)

Operating Expenses in Retail Grocery Stores: 1924, No. 52; 1923, No. 41; 1919, No. 18 ..... 50 cents each
No. 13. Management Problems in Retail Grocery Stores (1918) ..... 50 cents
No. 5. Expenses in Operating Retail Grocery Stores (1914) ..... 50 cents
No. 3. Operating Accounts for Retail Grocery Stores (revised edition — 1922) ..... 50 cents

GROCERY — WHOLESALE (*See also* CHAIN STORES)

No. 55. Cases on Merchandise Control in the Wholesale Grocery Business (1925) ..... (In cloth) $1.00
Operating Expenses in the Wholesale Grocery Business: 1923, No. 40; 1921, No. 30; 1919, No. 19. 50 cents each
No. 14. Methods of Paying Salesmen, and Operating Expenses in the Wholesale Grocery Business in 1918 50 cents
No. 9. Operating Expenses in the Wholesale Grocery Business (1916) ..... 50 cents
No. 8. Operating Accounts for Wholesale Grocers (revised edition — 1920) ..... 50 cents

GROCERY — MANUFACTURERS

No. 79. Marketing Expenses of Grocery Manufacturers for 1927 and 1928 ..... $2.00
No. 77. Marketing Expenses of Grocery Manufacturers for 1927 ..... $1.50
No. 69. Marketing Expense Classification for Grocery Manufacturers (1928) $1.50

HARDWARE — RETAIL

No. 21. Operating Expenses in Retail Hardware Stores in 1919 ..... 50 cents
No. 11. System of Operating Accounts for Hardware Retailers (1918) ..... 50 cents

JEWELRY — RETAIL

No. 76. Operating Results of Retail Jewelry Stores for 1927 ..... $1.50
No. 65. Operating Expenses of Retail Jewelry Stores in 1926 ..... $1.50
Corresponding Bulletins for earlier years: No. 58, 1925; No. 54, 1924; No. 47, 1923; No. 38, 1922; No. 32, 1921; No. 27, 1920; No. 23, 1919 ..... 50 cents each
No. 15. Operating Accounts for Retail Jewelry Stores (1919) ..... 50 cents

LABOR

No. 25. Labor Terminology (1921) ..... (In cloth) $1.00

PAINT AND VARNISH — WHOLESALE

No. 66. Operating Expenses in the Wholesale Paint and Varnish Business in 1926 ..... $1.50
No. 60. Preliminary Report on Operating Expenses in the Wholesale Paint and Varnish Business in 1925. 50 cents

PLUMBING AND HEATING SUPPLY — WHOLESALE

No. 72. Methods of Departmentizing Merchandise and Expense Figures for Plumbing and Heating Supply Wholesalers (1928) $1.00
No. 71. Operating Expenses of Plumbing and Heating Supply Wholesalers in the Central States in 1927 .... $1.50

PRIVATE SCHOOLS

No. 62. Operating Expenses of Private Schools for the Year 1925–26 ..... $1.00

PUBLIC UTILITIES

No. 68. Interstate Transmission of Power by Electric Light and Power Companies in 1926 ..... $2.00

SHOE — RETAIL (*See also* CHAIN STORES)

No. 59. Cases on Merchandise Control in Women's Shoe Departments of Department Stores (1926) $2.00
Operating Expenses in Retail Shoe Stores: 1923, No. 43; 1922, No. 36; 1921, No. 31; 1919, No. 20 ... 50 cents each
No. 10. Management Problems in Retail Shoe Stores (1913–1917) ..... 50 cents
No. 7. System of Stock-keeping for Retail Shoe Stores (1922) ..... 50 cents
No. 2. Operating Accounts for Retail Shoe Stores (revised edition — 1917) ..... 50 cents

SHOE — WHOLESALE

No. 6. System of Accounts for Shoe Wholesalers (1916) ..... 50 cents

STATIONERY AND OFFICE OUTFITTING — RETAIL

No. 80. Operating Results of Retail Stationers and Office Outfitters in 1928 ..... $2.00
No. 67. Operating Expenses of Retail Stationers and Office Outfitters in 1926 ..... $1.50

TEXTILES (*See also* COTTON)

No. 56. Distribution of Textiles (1926) ..... (In cloth) $3.50

WALL PAPER — WHOLESALE

No. 73. Operating Expenses of Wall Paper Wholesalers in 1927 ..... $1.50