J. Soc. Cosmet. Chem. **22** 211-229 (1971)

Sensory testing—a statistician's approach

W. A. PRIDMORE*

Presented on 22nd April 1970 at the Symposium on "Perfumery", organised by the British Society of Perfumers and the Society of Cosmetic Chemists of Great Britain, at Eastbourne, Sussex.

Synopsis—This paper shows ways in which the statistician can assist the cosmetic chemist in SENSORY TESTING by (1) employing his knowledge of PROBABILITY, (2) stressing the importance of a fuller understanding of the purposes for which the sensory tests are being used, and (3) stressing the importance of EXPERIMENTAL DESIGN in getting the full benefit from any set of tests.

It sets out the theory behind certain forms of tests of differences and preferences.

Introduction

This paper deals with some techniques devised by one group of statisticians to help the cosmetic chemist in the appraisal of new product formulations, in the development and modification of existing products, and in the routine control of raw materials and of production.

It is always difficult in any experimental or investigational process with which a statistician is associated to distinguish between that part of the plan which is directly attributable to the statistician's intervention and the basic experience, scientific knowledge and technical know-how contributed by the rest of the team. It is therefore to be understood that, whilst this paper will stress the statistical and probability aspects of the work which link together the various parts, most of the ideas have come from the statistician's technical colleagues, sometimes as a spontaneous contribution, sometimes in response to the statistician's awkward questions.

*Reckitt & Colman Products Ltd., Hull, Yorks.

The statistician can only work as a junior member of a team; he takes the ideas of others and helps them to put those ideas to work as efficiently as possible.

At the start, therefore, this statistician acknowledges that most of the good ideas presented have originated in the first place from other people; any nonsense is his own responsibility.

We shall be confining ourselves to that important but limited field of investigation in which the research chemist seeks to supplement his own judgement about the sensory characteristics of the products he is dealing with by calling upon the judgement of other judges or members of panels who are available at short notice in the laboratory area or accessible to it.

The statistician's role

A statistician may be asked to provide some analysis of results obtained by going to outside groups of potential users of products scattered throughout the country on a nationwide basis. Such market research tests, market placement tests, consumer panel tests, call them what you will, have an unavoidable tendency to be time consuming, costly and difficult to organise. The inevitable delays imposed by the problems of packaging for distribution over the whole country, of the distribution of product and the collection of reports are all difficult to accept at a time when commercial executives are breathing heavily down the formulator's neck. Some quick and simple approximations are absolutely necessary.

On the other hand, the statistician has known of the opposite dangers; the instant decision achieved by the chemist's own personal choice, which so often leads to expressions of acute surprise when larger scale tests fail to confirm the results from that sample of one. A very similar phenomenon concerns the use of the managing director's wife; she rarely constitutes a very typical market for the products in question.

In the first instance, the statistician may be required to advise on cheap and speedy versions of the national consumer panel of testers; in the second instance he may be required to provide some systematic substitute for the one or two judgements by the cosmetic chemist and his colleagues at the next laboratory bench. From whichever direction the approach is made, we tend to end up with a small-scale sensory testing panel, in some cases composed entirely of the non-technical office or factory staff immediately available on the work site, in other cases composed of specially recruited outside groups of people called together for special sessions to some con-

venient testing site, perhaps a local hall or a mobile caravan in a market place.

The most likely starting point for the statistician in this field is to be confronted with some single judgement by a research worker and be asked to support it with an experiment designed to convince all concerned that this is a valid statement of fact. Alternatively, some point of dissension between two judgements occurs, and the statistician is brought in to devise a crucial test.

The statistical line of attack is to request from the client full details of the experimental situation and to seek to devise a suitable and relevant testable hypothesis. For example, let us suppose that the research worker is interested in whether a difference in flavour exists between two batches of a peppermint oil for incorporation in a tooth preparation. He may be concerned to evaluate whether the addition of a bacteriostat to a cosmetic cream has caused a detectable change in its perfume. He may wish to evaluate the effect on colour of n-months' storage in plastic containers in comparison with storage in glass containers.

Theoretical probability considerations

All of these are clearly concerned with sensory difference testing. Sir Ronald Fisher (1) set out in some detail the statistical principles involved in setting such a testable hypothesis. He described in a classical reference "a lady who declares that, by tasting a cup of tea made with milk, she can discriminate whether the milk or the tea infusion was first added to the cup". The hypothesis to be tested was that she was unable to discriminate between the two forms of tea, and that her identification was, therefore, purely at random. The statistician calls this the null hypothesis.

The experiment devised by Fisher was to offer the good lady eight cups of tea in turn, four being mixed in one way and four in the other. These were presented to her in a random order and she had to taste each of the eight and identify whether tea had been added to the milk or the milk added to the tea. On the null hypothesis (that is that her identification was made purely at random), the probability of her making a completely correct set of eight identifications, assuming she knew that there were four of each, would be 1 in 70.

It is the statistician's approach to such matters to assume that, should such an unlikely event occur, then its occurrence should be taken to be evidence that the null hypothesis is not true. Thus, if Fisher's lady correctly identified all eight cups of tea, then she was not choosing at random.

In sensory testing using taste, it has long been customary to use tests of this form. The most widely used is the Triangle Test in which panel members are offered three samples, two of one kind and one of another, and are requested to identify the odd sample (2–6). In such a test, which has to be carefully drawn up and executed to present every possible ordering of the samples in a balanced experimental design, so that the two samples appear as odd samples an equal number of times and in such ordering (i.e. the orderings ABB, BAB, BBA, AAB, ABA, BAA, but in random sequences), the panel members have a 1 in 3 probability of picking the odd sample by chance even if they have no discriminating power.

It is therefore necessary to rely on panels of judges to establish whether there is a distinguishable difference between two samples; if three people correctly perform a triangle test, the probability of them doing so by pure chance is but $(\frac{1}{3})^3$ or 1 in 27. Tables are available for testing the statistical significance of less conclusive results. Thus, if out of a panel of 20 members, 11 correctly identify the odd samples instead of the seven which would be expected on the null hypothesis, then the probability of so doing is less than the conventional 1 in 20 level used so often by statisticians and others when they can think of no valid reason for choosing any other level. A simple table may be found in Ostle (6).

A similar procedure has been used by Harries (7) in situations involving the testing of foodstuffs. Here the panel members are offered three samples of one kind and two of the other (in random order) and the probability of a correct classification assuming a null hypothesis of chance selection is 1 in 10. Clearly such a test procedure will be more sensitive than the triangle test, and fewer panel members will be required to establish that a given sensory difference is distinguishable.

Indeed, the principle involved may be generalised to an n+m test, in

Table I.

Probability of correct discrimination between samples on assumption of random selection.

		n					
		1	2	3	4	5	6
m	1	(1.000)	0.333	0.250	0.200	0.167	0.143
	2		(0.333)	0.100	0.067	0.048	0.036
	3			(0.100)	0.029	0.018	0.012
	4				(0.029)	0.008	0.005
	5					(0.008)	0.002
	6						(0.002)

which a total set of n+m samples, n being of one kind and m being of the other, are given to the panel members to sort into the two categories. *Table I* presents the probabilities for such tests.

Table I only records half of the probabilities, but it is to be noted that $(n+m) \equiv (m+n)$. The diagonal entries printed in brackets give the probabilities for those cases where the total set of samples is divided into two equal groups; they also refer to the case when the panel member is not required to identify either of those groups in any way. Should the panel member be asked to identify one set as being (for example) the stronger perfume, then those probabilities should be halved.

Efficiency of designs for assessing whether detectable differences exist may be assessed by finding the lowest probability for a particular total number of samples. The most efficient designs for discrimination are to be found when $n-m=1$; the reader may check this for himself from *Table I*. It should also be noted that the probability associated with a test in which $n=m$ is the same as that for $n+m$ where $m=n-1$, e.g. the probability of discrimination on the null hypothesis for a 2+2 test is $\frac{1}{3}$, which is exactly the same as for a 2+1 test.

However, it is to be noted that the conduct of a test in which $n \neq m$ is complicated by the fact that the experimenter has to make a decision about which of the two preparations is allocated to n and which to m. This is not a trivial decision because experience shows that in these two cases, the probability of discrimination may not be the same. In triangle (2+1) tests of flavour, for example, in which peppermint oils are being compared, it has been shown that it is easier to detect a stronger flavour if it is being compared with two weaker flavours than vice-versa. For this reason all possible orderings should be tested; by far the simplest way of achieving this is to give the panel member equal numbers of each sample to test in whatever order he or she finds most simple to discriminate.

Moreover, as we shall see, the testing procedure is frequently extended to involve questions of preference; under such circumstances it can also be observed that the unbalanced designs (where $n \neq m$) appear to bias or at least to have an effect upon the preference judgements; preferences in some cases appear to go in favour of the larger number, and in other cases in favour of the smaller number by margins that cannot be due to chance. It appears desirable to eliminate this complication by making $n=m$.

Let us therefore examine more closely those balanced test designs. The 1+1 test is clearly a nonsense; we can always correctly divide up A from B without this telling us anything, hence the probability of correctly dividing

by chance is 1.0, i.e. complete certainty. However, with 2+2, we have one chance in three of correctly putting the two A's together and the two B's together. It will, of course, be noted that with the symmetrical designs, it does not matter which of the two sets we nominate as A and which we nominate as B; if there is one A amongst the B's, then there will clearly be one B amongst the A's at the end of the sorting procedure. *Table II* demonstrates the procedure.

Table II.

Balanced sorting designs (n=m)
Probabilities of errors on random sorting hypothesis

		n					
		1	2	3	4	5	6
Number of errors (e)	0	1.000	0.333	0.100	0.029	0.008	0.002
	1	(0)	0.667	0.900	0.457	0.198	0.078
	2				0.514	0.794	0.487
	3						0.433

From this table it can be seen that a 4+4 design is the smallest sized design which provides for an individual judge to show significant discrimination at a probability level of 1 in 20 (p=0.029). However, if one goes up as far as the 6+6 design, not only can we demonstrate that complete correct division with no errors is most unlikely on the hypothesis of random selection (p=0.002), but it is also possible to consider a second grade of response in which there is only one error (i.e. five identical samples and one misplaced sample in a set of six and this still has a low probability of occurrence (p=0.078) on the assumption of pure random choice.

As we shall go on to show, the 6+6 form of test has a number of convenient features for the making of odour comparisons, and these are largely concerned with the evidence about the detectability of differences by individual panel members.

However, the statistician has to examine not only the probability of the panel member making a correct selection entirely by chance (known to the statistician as a Type 1 error) but also the probability of the panel member making an incorrect selection even though he is able to discriminate (known as a Type 2 error). However, in order to do this, it is necessary to consider the mechanism by which the panel member is thought to make his selection.

The most reasonable basic concept is one proposed by Gridgeman (8);

he suggests that there is a probability (p) of the panel member making a correct sensory perception of the nature of the sample. Thus in an n+n test, there is a p^n chance of correctly identifying the group of n; using a binomial distribution, it is possible to compute probabilities of correctly identifying any number of samples up to n. However, when less than n items are correctly identified, the model has also provision for the possibility of further "correct" selections by chance.

In the terminology used earlier in this paper, which differs slightly from that used by Gridgeman:

$n+m=t$ (For the cases in *Table II*, $n=m=\frac{t}{2}$)

p= probability of making a single correct identification (i.e. not by pure random selection)

$q=1-p$

x=number of correct identifications

r=number of correct allocations (including random selections in addition)

Then $t>m\geqslant r\geqslant x\geqslant 0$

$$P\,(r;m)=\frac{m!}{(t-2m+r)!}\left[\frac{(t-m)!}{(m-r)!}\right]\cdot\sum_{x=0}^{x=r}\frac{(m-x)!}{(t-x)!x!}\,p^{x}\,q^{m-x}$$

Using this approach, we are able to calculate a series of probabilities assuming different levels of discrimination on the part of panel members.

Table III.

Probability of completed correct allocation for given probability of basic discrimination.

Nature of test	Response	0	0.1	0.5	0.9	1.0
2+2 test	2 correct	0.330 0	0.340 0	0.500 0	0.873 4	1.000 0
3+3 test	3 correct	0.100 0	0.104 8	0.268 8	0.792 6	1.000 0
4+4 test	4 correct	0.028 6	0.031 1	0.146 4	0.717 7	1.000 0
5+5 test	5 correct	0.007 9	0.009 1	0.079 2	0.648 8	1.000 0
6+6 test	6 correct	0.002 2	0.002 8	0.042 4	0.585 7	1.000 0
	5 & 6 correct	0.077 9	0.086 7	0.325 6	0.934 8	1.000 0

From *Table III* then we show not only the probability of an individual's correct discrimination, assuming the null hypothesis of pure random selection, but also the probabilities of correct discrimination assuming

various alternative hypotheses; this enables us to consider the power of the test in the statistician's language. However, when we start to deal with groups of people collected together in a panel then we start at once to meet practical difficulties; it is a well-known fact of life that panel members have differing basic nasal sensitivities; it is not enough to consider a large group of people all with the same sensitivity. This means that we can expect to find in a panel of 30 people several who are extremely sensitive to the difference in question and a considerable proportion slightly aware of the same difference. The complications which beset the statistician when he gets away from the simple null hypothesis are tied up with this problem of a distribution of sensitivities and there appears to be little practical advantage in pursuing this matter very much further, except to note that the 6+6 test enables us to make quite useful statements about each individual member of the panel, instead of as with the smaller tests (e.g. 2+2, 3+3) where we can only talk meaningfully about the panel as a whole.

The following table (*Table IV*) gives details for three conventional panel sizes, the numbers of positive results required to show the existence of meaningful differences from the null hypothesis using the 6+6 test design.

Table IV.

6+6 Test design (Tests of significance)

Panel size	Number of panel scoring 6/6	Number of panel members scoring 5/6		
		Fairly sure (P<0.05)	Sure (P<0.01)	Almost certain (P<0.001)
10	0 combined with	4	4	6
	1 combined with	0	2	3
	2 or more *almost certain* under all circumstances			
20	0 combined with	5	6	8
	1 combined with	2	4	6
	2 or more *almost certain* under all circumstances			
30	0 combined with	7	8	10
	1 combined with	3	5	7
	2 combined with	0	0	4
	3 or more *almost certain* under all circumstances			

Some practical considerations with N+M test designs

This paper has discussed at some length the probability and statistical

aspects of these sensory testing procedures. To recapitulate, the procedure is as follows:–the panel member is confronted with n samples of one kind and m samples of the other which are coded in a randomly allocated numerical sequence; there should be no other detectable difference except the particular characteristic under test. The panel member is then requested to divide up the two sorts of sample into their appropriate groups using the relevant sense. If n+m is large enough, it may be possible to judge whether each panel member is a discriminator or not. If, on the other hand, the value of n+m is too small, it may not be possible to categorise each panel member, but it will still be possible, by using a large enough panel, to declare whether there is any evidence of a distinguishable difference from the results of the panel as a whole.

The essential elements of the test are that panel members can physically sort the unidentified samples using the sense to be tested; if this sorting agrees with the experimenter's undisclosed coding, then we can demonstrate a discriminable difference.

Why do such tests?

As we have already suggested, the purpose of the test may simply be to discover whether a detectable difference exists between two samples. However, it frequently happens that the basic requirement is one of establishing preference between samples; even where the primary concern is one of difference, we may still wish to know which of two different samples is the more acceptable.

There is an extension to this line of argument; it is to be taken as axiomatic that unless people can distinguish between two samples, they cannot validly express a preference for one rather than another. "I cannot tell the difference between these two samples, but I prefer that one" is not an acceptable statement from a panel member. One approach to this has been to select in advance a specialist group of panel members who specialise in distinguishing between particular types of products, or between certain brands within a product group. It is, however, a serious problem with any organisation dealing with any appreciable number of product fields that it would be necessary to set up a very large number of such specialist panels. What is more, these would require setting up in advance of every new piece of investigation, and this could be a severe limitation on the speed with which research is conducted.

However, for those fields of investigation for which an n+m test of sufficient size can be used (e.g. 4+4 and over), it is possible to carry out a

form of instant panel selection. By requiring a series of panel members to carry out such an n+m test, and then asking them to say which group of samples they prefer, it is possible to pick out those individuals who can discriminate and only taking note of their preferences. In this way it is possible to pick out an expert panel and carry out a preference test all in the one operation; indeed by this means it is possible to eliminate the preferences of those panel members who on the testing occasion in question were unwilling or unable (perhaps owing to coryza or catarrh) to discriminate.

For this purpose, particularly for testing of perfume or smell, the 6+6 test has proved admirably suited; panel members find themselves well able to assess twelve samples, the design is balanced and so bias is not introduced into the preference assessment, and the design provides for two grades of judgement, 6/6 or completely correct judgement (with a probability on a random null hypothesis, of 0.002), and 5/6, or a judgement with only one error, which still has a probability of less than 0.08.

This, then, is a device for selecting suitable panel members; one obtains first of all, an indication of the magnitude of differences existing between the two samples under test; one then proceeds to assess the acceptability of that difference by examining the preferences of those panel members who demonstrate that they are able to discriminate. These constitute a ready made specialist panel, whose ability to discriminate is assessed virtually simultaneously with the obtaining of their preference judgement.

What panels should be used?

The selection of judges, subjects, panel members (these names are virtually synonymous) is dependent, as always, upon the objectives for which the whole experiment is to be designed. The n+m test may be used to establish whether the view of a particular research worker about the existence of some sensory difference can be confirmed. In this instance, the research worker himself would be checked in a double blind trial to see if he or she could detect the difference; using the 6+6 test, the likelihood of any correct allocation of the samples being due to chance may be regarded as negligible. On the other hand, it is equally clear from *Table III* that the single observer, even though having a 9/10 probability of spotting the difference in question in a single observation, has still a very appreciable probability of not getting all his allocations correct (P of 6/6 is still only 0.59); however, it is an interesting phenomenon that the research worker in these circumstances is so taken aback at the thought that he is not 100%

perfect (and demonstrably so!) that he is willing to concede that perhaps the difference is not as great as was hitherto thought. It should be noted from *Table III*, however, that with the same 9/10 probability on the single identification, the likelihood of the single panel member getting 5 out of 6 in a 6+6 test is as high as 0.935.

At this point, it is appropriate to make use of wider panels of non-specialist individuals, preferably having nothing to do with the technical aspects of the product under consideration. It will be possible to establish then whether there are generally detectable differences present; it is usual to find that only a proportion of the panel can clearly discriminate. The general finding is that panel members are idiosyncratic in the differences they can detect; some panel members can detect certain differences and different sets of panel members are good at other differences. Yet again there are small groups who appear to be good at a very wide range of differences. A hypothesis which would be open to investigation is that these groupings may be genetically determined and could throw light on the mechanisms of the sense of smell.

How should the samples be presented?

It is clearly of the essence of the logic of the tests described that the differences between the samples should be made manifest only by the sense under test. The test procedure lays down that if the panel member sorts the samples into the correct groups, then there is evidence that the choice was by something other than random selection. It is, however, only by careful attention to the actual mechanism of presenting the samples to the subject that we can be in a position to assume that the only alternative basis for the subject's choice is the use of the relevant test.

It we are conducting a test on taste, the food technologist will wish to make sure that we cannot tell Stork from butter by its different colour; with an odour test of bath crystals, we shall wish to disguise colour differences in order to be sure that discrimination can only be made by the sense of smell; in an auditory test of violins (9), we shall wish to conceal the actual Stradivarius from the audience panel by a screen.

In the toiletries situation, where we are concerned with perfumes, we find it relatively easy to make comparison between alternative versions of our own products; it becomes more difficult if we wish to compare our own product with a competitor. The use of amber-glass jars conceals very minor variations in appearance, of the kind which are sometimes detectable in full light between powders perfumed with different perfume ingredients.

Much more difficult to conceal are the differences between clear liquids and cloudy suspensions which can arise in making smell comparisons between alternative formulations.

In these circumstances it is sometimes necessary to conceal the actual surface of the layer of product by a wire gauze baffle, or a layer of grinding rods or some such inert material. As a final resort, the use of special dim or coloured lighting or even the wearing of dark goggles is effective. If there is any doubt whatsoever, a preliminary test could be to require the subjects to sort the samples *without* using the sense (taste, smell, etc.) which is really to be tested; if the subjects succeed, then the conditions for the test are not appropriate.

For the statistician it should not be necessary to draw attention to the fact that the samples under test should be allocated to their code numbers at random. This was forcibly brought home to one statistician who lapsed into a convenient convention of always labelling one group of samples in a 6+6 test with odd numbers and the other with even numbers. This was convenient because the incorrect items in the sorted groups were at once obvious. However, when the test organisers carried out a test in which all twelve samples were poured out of a single bottle of bath additive, it was found that a significant proportion of the panel (at a probability quite beyond the accepted extreme 1 in 1 000 level) were dividing the samples up into odds and evens. It was therefore established that random numbering of samples under test was the only acceptable procedure.

Finally, it must be stressed that the form of presentation of the sample to the subject needs very careful consideration. Sensory tests of the kind we are discussing can only handle one aspect of the product at a time; we can, for example, carry out tests of the neat smell of a product in the jar, a face cream, a deodorant stick or what you will; we may test a shampoo when diluted down in water, or a bath preparation when steaming in hot water, or the residual smell left after using a hair spray, but whatever we test, we must take note of the fact that we are only considering and testing one particular facet of the organoleptic characteristics of the product. One of the more unwelcome tasks of the statistician is to have to draw attention to his clients that they may not generalise from such limited tests, say of the odour of a bath powder in the dry state to the perfume acceptability of the product in general. Too often it proves to be only too convenient to accept the easily obtained answer instead of the relevant one; unfortunately we so often never get enough information to know how misled we have been.

The subject's response

We have so far considered these tests from the point of view of the test organiser; the statistician has to be aware of the fact that it is of no avail designing elaborate procedures and methods for their analysis if it is not possible to obtain enough experimental material – in this instance, people willing and able to act as subjects.

It will be recalled that although the discussion has so far been concerned with looking for differences, we have referred to the use of these techniques to assist with preference assessments. Working on the assumption that only those people who are aware of a difference have any valid basis for having a preference, it is possible to use the higher order n+n tests as a means of jointly assessing the ability to detect a difference and, having done this, to only take note of the preferences of those individuals who have made the correct responses.

We ourselves make extensive use of the 6+6 test for the purpose of assessing odour; as we have seen, this effectively permits us to identify the individual panel members who can discriminate, and who can discriminate on that particular occasion. In short, we pick a specialist panel for the job in hand, and we pick that panel for sensitivity at the very time that we are interested in obtaining their choices. Thus, if they are temporarily suffering from catarrh and cannot use their sense of smell, then they automatically exclude themselves on that occasion only, without prejudice to their further selection on another occasion. Moreover, if the panel member is not well motivated, and does not adequately smell the samples, then he or she will likewise fail to discriminate and therefore will effectively be excluded from the panel. This gets over the whole problem of a rapidly changing pattern of testing; if pre-picked panels were used, the greater part of the time would be occupied in selecting new panels rather than using the existing ones. By the use of the techniques just described, we select and use simultaneously.

A typical set of results using a panel size of 30 looks as follows – the

Table V.

Result of a 6+6 test (R. & C. O.P. test 6 200)

	Prefer A	No preference	Prefer B	
6 correct	3	–	–	10%
5 correct	6	5	2	43%
4 correct	4	5	1	33%
3 correct	–	4	–	14%
		(n=30)		

figures are taken from a test on an alternative de-naturant used for the alcohol in an aftershave.

From this we can deduce the following: of the total panel of 30, 16 (53%) were able to discriminate (got six or five correct, significant at $P < 0.001$, see *Table IV*), and of those discriminating, nine preferred A and two preferred B (with five declaring no preference) showing a clear preference for A.

Long experience shows that panel members not merely do these tests but enjoy doing them. The task gives an intellectual challenge and there is no problem of motivation; panel members ask eagerly if they have "got the test right", and this it is possible to tell them as an incentive to do better. A straight preference comparison between two samples has no "right" or "wrong" answer, and the subject who asks whether his or her answer is "right" in a paired comparison is under a misapprehension about the purpose of preference questions of this kind. But the challenge of the n+m type of test, with its right and wrong answers which may be communicated to the subject without invalidating further testing, leads to considerable enthusiasm and maintains a high degree of motivation over long years of testing.

One question which is frequently asked about this form of test concerns sensory fatigue; it is suggested that sensory fatigue rules out the possibility of a subject smelling as many as twelve samples and correctly distinguishing between them. Our experience is otherwise; in the field of odour we have no difficulty in getting subjects to distinguish between sets of twelve with quite trivial differences between them. Indeed, for an investigation of this very point, we arranged for a panel of some 33 members to repeat the same test four times in succession (forty-eight jars smelled altogether) and the rate of discrimination (52%, 58%, 50%, 59%) remained effectively constant over the four tests.

On the other hand, there are limitations, not so much of physical sensory fatigue, but of ability to get rid of one substance before sensing the next which limits the ability to taste long series of samples. 2+1 (triangle tests) are usually the upper limit to the size of test that can be offered to the taste panellist. Pungent tastes which linger in the mouth and deaden the taste buds are difficult to classify; whether they represent true sensory fatigue is perhaps arguable.

It would be wrong to suggest that sensory fatigue was not present in the larger scale 6+6 type testing of smells, however constant the response remained over time. It would be arguable that it makes discrimination of

small differences more effective, if its effect is to cancel out the common elements of two odours, and merely leave those portions of the odours which were not common to both samples more obvious. There is some support for this hypothesis in the finding that the 6+6 test is relatively insensitive to large differences in perfume concentration (where the ratio may be of the order of 100:150) in cases where we are talking about identical perfumes at different levels, but is extremely sensitive to quite small contaminant traces of extraneous odour.

However, it is arguable that if sensory fatigue does enter into the judgements of panel members assessing smells under these test conditions, then this is entirely appropriate since, particularly with personal cosmetic products; it is the smell after continued exposure which requires to be judged.

Further developments

This document has confined itself to discussing in some detail the statistician's approach to certain forms of sensory testing in which two different samples are compared together.

Much work has now been carried out on an extension to these techniques in the odour testing field in which many more different products are used (up to 10 or 12) and in which the subject is given the opportunity to create whatever groups he or she thinks appropriate. This approach has proved exceedingly valuable in the quality control field where there are a number of batch samples to compare with some form of control sample; this particular application is built upon the structure of the 6+6 test described in the earlier part of this report, which is used to provided known and measured differences to be inserted in amongst the set of smells to be evaluated.

The procedure is to take the series of batch samples and to bulk them together; having done this the bulk sample is divided in half, and one half has a known amount of a contaminating substance mixed in with it. This difference is checked by a 6+6 test, which is expected to yield a difference of about 30% scoring six and five correct. This level was hit upon empirically as the result of long testing on a variety of different formulae changes of greater or lesser degree; 30% was established to be the kind of test difference found between samples available in the retail trade at the end of two years at a time when no complaints whatsoever about smell were being received.

This and similar arguments led to the establishment of a 30% discrimina-

tion level; it is now interesting to see from *Table III* that this level corresponds to a basic p of 0.5; in other words, this is the kind of level of detectability at which a panel member is as likely to say a difference exists as often as not whenever he judges it. This seems to be a very logical sort of level to take for the base line kind of differences below which one does not want to bother to know.

Having built this known difference into the test set, then the operation calls for panel members to group the samples into as many different groups as they can detect. Any samples which are grouped together more frequently than the difference between the Bulk and the Contaminated Bulk (to use the terminology used in the test situation) are then indistinguishable; the test samples should be indistinguishable from each other and a control sample.

This same technique is combined with some multivariate statistical analyses employing a mathematical procedure known as the method of Principal Components to enable meaningful interpretation to be put on the way in which the samples are grouped together. It is not proposed to discuss this further in this paper except to indicate that the techniques we have described have considerable potential for further use and development.

Conclusion

This paper has been written to show that the close collaboration of the statistician and the cosmetic chemist in this difficult field of sensory assessment can be of considerable mutual benefit. In particular, I would lay stress on the following points:–

(1) This is a field in which the concepts of probability are of vital importance – the probability of making choices regardless of any discriminating power as compared with the probability to be found when making discrimination.

(2) This is a field in which it is very easy to lose sight of the essential purposes of the tests unless great care is exercised by both statistician and the cosmetic chemist. In particular, it is too easy to assume that too much has been proved by a very limited test simply because it makes life complicated to believe otherwise.

(3) This is a field in which the statistician can use the tools of his trade to get out a lot of hidden information, provided he can suggest the experimental designs before the work is done, not afterwards.

(Received: 7th February 1970)

REFERENCES

(1) Fisher, R. A. *Design of experiments* 6th Edition. 11 (1935) (Oliver & Boyd, Edinburgh & London).
(2) Bengtsson, K. and Helm, E. Principles of taste testing. *Wallerstein Lab. Comm.* **9** 171 (1946).
(3) Lockhart, E. E. Binomial systems and organoleptic analysis. *Food Technol.* **5** 428 (1951).
(4) Green, M. W. A note on the use of triangle design in taste testing of pharmaceutical preparations. *J. Am. Pharm. Assoc. Sci. Ed.* **44** 380 (1955).
(5) Fourman, V. G. Taste panels for pharmaceutical flavors. *Drug Cosmetic Ind.* **77** 762 (1955).
(6) Ostle, B. *Statistics in research.* 58 (1954) (Iowa State College Press, Iowa).
(7) Harries, J. M. Sensory tests and consumer acceptance. *J. Sci. Food Agric.* **4** 477 (1953).
(8) Gridgeman, N. T. Sensory item sorting. *Biometrics,* **15** 298 (1959).
(9) Richardson, E. G. The science of orchestral instruments: Some recent work. *Discovery* **14** 87 (1953).

DISCUSSION

MR. A. ELLIOTT: At this Symposium three distinguished speakers have recommended three different methods of selecting perfumes for new cosmetic products. Mr. Erni (10) suggested it should be the perfumer's responsibility to select the perfume, Mr. Landon (11) suggested the marketing manager should make the final choice, and finally you suggested that the statistician, with the aid of panel tests, should choose the perfume.

The suppliers of perfume compounds are anxious to point out that a distinctive perfume will probably only be liked by about 15% of the population and yet still be a commercial success, whereas a perfume which is generally acceptable is not distinctive and hence will not sell the product.

In view of this, can you tell me of a method of panel testing which will select a distinctive, rather than generally acceptable, perfume for new products?

THE LECTURER: It is not my view that the statistician should have anything to do with the actual selection of the perfume. It should be the potential customer, when all is said and done—and all that the statistician can offer to do is to help both the cosmetician and the marketing man to find the best way of obtaining this information about the potential customer.

The question of how one goes about looking for the proportion of people that are going to find the perfume acceptable is very involved and it depends to what group you are aiming your product. It is not a statistician's function to answer (only to ask), and what he can do is simply to draw up rules for saying—if you are making decisions like this (which are, after all, commercial decisions or, possibly, perfumery decisions) what kind of samples you have to draw, what kind of numbers of people you are going to have to interview before you have some reasonable assurance of having met that target. Without talking about specific cases (and I do not think there is much point in doing that here), this is a question of laying down specifications.

MR. C. SEBLEY: How much rechecking takes place with your panel, not so much from the point of view of sensory fatigue, but from a "preference" angle?

(10) Erni, M. Evolution in perfumery. Presented at the Symposium on "Perfumery" at Eastbourne, on 21st April 1970.
(11) Landon, M. F. The application of perfume to cosmetics and toiletries. Presented at the Symposium on "Perfumery" at Eastbourne, on 20th April 1970.

in assessing general acceptability would be very much governed by many other things, some of which, like brand image and promise, are controlled from the commercial side. We have confirmation that the basic characteristics of the smell acceptability have remained similar in in-use situations, to the answers that we obtained in the small-scale testing such as described. What does not always follow is that this has the direct effect on the general acceptability of the product that one hoped for; one finds that because the final consumer is using the product in a total use situation (as distinct from just smelling it), we have a halo effect operating in both directions. Perfume acceptability may be having a "halo effect" upon the "acceptability in use" of the product but, of course, also vice versa.

Therefore the fact that on some occasions we get discrepancies between the overall in-use acceptability and the perfume test is something that we have to live with.

MR. P. MOXEY: I understand from what you said that you expunge people if they are not necessarily able to give you what one might call a distinct answer. What concerns me is that a great many products are bought in this business on the strength of the fragrance, and I can understand why you want to get your statistics right to start with. I wonder if statistically you are not building a false image into your preliminary testing.

THE LECTURER: What you say is very true. First of all I would say that we do not exclude people from the panel except in the sense that we only look at their preferences if they show evidence of being able to discriminate. They do come back the next time and are again put into the test because we do not exclude the people who last time were unable to discriminate. We find that you have a variable level not only of acceptability but also of discrimination – and this is something that happens in real life as you yourself have indicated. Sometimes people are strongly motivated about a product's smell, sometimes they could not care less whether it smells nice or not. What we are saying is that, in so far as they do make their choice in terms of smell, we wish to have the smell most acceptable to the majority of people. If they are not choosing on the basis of smell, that is just too bad. We can not do anything about that. What we try to make sure is that on the occasions that people are choosing on the basis of smell, this product will be as acceptable as we know how to make it.