The Flavor Problem of Soybean Oil
II. Organoleptic Evaluation

HELEN A. MOSER, CAROL M. JAEGER, J. C. COWAN, and H. J. DUTTON
Northern Regional Research Laboratory,[1] Peoria, Illinois

PROGRESS on the problem of flavor stability in soybean oil has been limited to a large extent by means available for evaluating the oils, and as yet no substitute has been found for the human sense of taste and smell in the study of flavor problems. While the hope of all research workers in the field is to replace the erratic human senses with objective physical and chemical analytical methods, it must be remembered that the ultimate evaluation of flavor is subjective. As long as human beings are the final judges of flavor, organoleptic evaluation will probably be required in flavor problems.

Panels for organoleptic evaluation have been organized for two general purposes, (a) as a research analytical method, and (b) as an index of consumer acceptance (1). As a research analytical tool for the measurement of the effect of variations in processing of food products, the selection, training, sensitivity, and consistency of individuals comprising a panel are of paramount importance. As an index of consumer acceptance, complete randomness in the choice of panel members and normal variations in prejudices and sensitivities are desirable.

The character of research on the flavor problem of soybean oil at the Northern Regional Research Laboratory has required the organization of the analytical type of taste panel. The function of this panel is to measure the effect of changes in refining methods upon flavor stability. The procedure of organoleptic evaluation used has been evolved over a period of several years and is constantly undergoing revision as new problems arise or more accuracy is required. The present refinements of this method lie primarily in the mode of preparation and presentation of samples and in the application of statistical methods to the selection and control of the panel and the evaluation of its results. Although in no sense at a final stage, the procedure herein described has given consistent results.

Selection and Training of the Panel. In the selection of the first tasting panel, preliminary acuity tests were given to 35 persons. These tests, patterned somewhat after the procedure outlined by the Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture (2, 3), for conducting taste and smell tests, were designed to measure thresholds and to compare the individuals' capacities to detect small differences in concentration. From this first test, 21 individuals were eliminated who had very high thresholds or made incorrect identifications. To the remaining 14 persons, a second series of tests was given as a check on their performance. In the final selection of the panel, however, certain subjective considerations, such as past experience on organoleptic panels, interest in the oil problem in general, and the desire to participate were involved so that the individuals chosen included others besides those who had low thresholds and showed acute sensitivity to the primary tastes.

Eight persons were chosen to be on the regular panel, and five others were designated as alternates. To begin with, a series of oils, previously rated by industrial experts, were given to the entire group as a method of education. During the following year many oils were given in many combinations, with the tasters trying to standardize their numerical and descriptive scores. There were changes in the personnel of the panel and not all of the new members were given the acuity tests, but all did gain familiarity with the technique of tasting oils. When the present approach to the flavor problem was begun, there were 11 members who had had this experience in tasting oils, and who had shown consistency and sensitivity by tests evaluated through the use of the statistical technique known as "Analysis of Variance" (4).

The designing of an experiment is a major consideration in the conduct of organoleptic evaluation. We shall mention, for illustration, some of the factors that had to be taken into consideration in setting up a recent experiment cited later in this paper. The evaluation by a panel member on a given sample may vary from day to day and, in addition, the evaluation will vary somewhat depending on the sample with which it is compared. It is recognized in the tasting of oils that the senses become saturated rapidly. For the most precise results it has been found best to present only two samples at a time, i.e., the "paired sample" technique. Each pair answers one specific question and evaluates a single variant of the processing. By the technique of paired samples the difference between pairs, which is of primary concern, is evaluated. The comparison is made by use of the "t" test. With this test, it is possible to determine whether an observed difference is significant. Detailed information concerning the mechanics of applying this test is included in Appendix A.

Conduct of the Panel. In the organoleptic evaluation of oils every effort is made to eliminate variables which might contribute to error in the scoring of samples. The taste panel meets in a well-lighted, air-conditioned room. Thus the air temperature is constant the year around, and there is a minimum of foreign odors. Each panel member is assigned a booth where, in a comfortable sitting position, he may evaluate the samples presented. This arrangement minimizes distractions and discourages the tendency to make audible remarks or otherwise convey impressions. Experience has shown that quietness, smoothness of presentation, orderliness, and regularity contribute to more accurate evaluations.

Odors and flavors of oils are more readily detected if the oils are warmed. Consistency demands that they be evaluated at the same temperature at each

[1] One of the laboratories of the Bureau of Agricultural and Industrial Chemistry, Agricultural Research Administration, U. S. Department of Agriculture.

tasting. To meet this requirement a table of special design was constructed, with its top consisting of an electrically heated aluminum plate. Small aluminum blocks for holding as many as six 50-ml. beakers were also constructed. The table top holds 12 of these blocks, one for each panel member. The blocks can be heated to 55° C. in one-half hour.

Samples are presented "blind" to the panel and usually in pairs. Each sample consists of 7.5 ml. of oil (or shortening) in a 50-ml. beaker covered with a watch glass. A paper cup filled with water at body temperature is provided for rinsing the mouth between samples. No sample is swallowed regardless of its nature. The score sheet completes the booth "set-up." On entering the room, the panel member is immediately presented with his samples and scoring begins. Figures 1 and 2 show the booth arrangement and the panel in operation.

FIG. 1. Panel members tasting.

A uniform procedure for tasting has been evolved. Beakers are numbered "1" and "2" and the members sniff the contents of the beakers in that order. In general, the oils having the least odor will also have the least flavor and therefore, tasting should be in order of increasing odor to avoid dulling the taste. After the samples are sniffed, the odor scores and descriptions are recorded. The oils are then taken into the mouth and allowed to reach the back as well as the front surfaces of the tongue. The sample is held in the mouth for about a half minute and then discarded in the paper cup provided for that purpose. The flavor score, as well as a description of the predominating flavor, is recorded on the score sheet.

The score sheet given in Figure 3 is used to record the intensity and quality of flavor. The panel member has a 10-degree numerical range for scoring each sample. Space is provided opposite the intensity for recording a word description of the odor and flavor. Any scoring system which may be devised is necessarily arbitrary and necessarily a compromise. However, after testing a number of variations in scoring methods, the present one was settled upon as meeting most adequately the requirements of statistical analysis and the preferences of the tasters. Bland oils are rated 10 in this system; mild buttery and other acceptable flavors are rated 7 and 8; painty, rancid, and grassy flavors fall in the range from 1 to 6 depending upon their intensity.

FIG. 2. Comparing results after samples have been evaluated.

On completion of tasting and scoring the samples, panel members are permitted to discuss and compare evaluations. Generally the description of the samples is given to the panel members. "Rewards" in the form of cookies made from shortenings of "acceptable" quality are provided which help to remove the taste of badly "reverted" samples.

The successful conduct of a taste panel is frequently as much a matter of human relations as it is a scientific problem. Individuals on the panel must have a keen interest in their tasting ability and these feelings must be sustained. Panel euphoria is thus another of the less tangible but yet real variables encountered in taste-panel operation. Frequently, individuals serving on the panel do so at the sacrifice of time from their own equally important research problems. The interest of panel members is sustained only by sharing with them the research developments. Informal conferences are held periodically to present results, to discuss plans for further experiments, and to inform the members as to how their individual tasting scores compare with that of the panel average.

Statistical Evaluation of the Panel. To obtain measures of the performance of the panel and the individuals comprising the panel, two methods of analysis were used. One, the "control chart" method, measures the reproducibility of the individual's scoring on a single oil; the other, a correlation and regression method, measures the ability to distinguish between different oils. For the "control" method one oil was given seven times in various combinations. The scores assigned by individuals to this oil were used to prepare "control" type charts (5, 6) of the average values and of the standard deviations. By this method the individual's average values and standard deviations are shown in addition to the average score for the panel and the limits. These limits are used to express statistical stability of the kind that may be expected in random samples from homogeneous material. These control charts are shown in Figures 4 and 5. The discontinuous lines for the limits result from the fact that not all the panel members tasted all seven combinations, and allowance has been made for the unequal number of tastings.

From the chart on averages (Fig. 4), it is apparent that only two of the members were outside the limits.

NM—281

NAME *John Doe*

ORGANOLEPTIC TESTS FOR RUN *42*

DATE *3-1-46*

Please describe the predominating odor and flavor of each sample in the space opposite the proper intensity. The lower figure in each bracket indicates the more intense odor and flavor.

		Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
		O	F	O	F	O	F	O	F	O	F
Good	10	*Bland*									
	9		*V. sl. buttery*								
Less desirable but acceptable	8										
	7										
Objectionable	6			*Painty*							
	5										
Unpleasant	4				*Painty*						
	3										
Repulsive	2										
	1										

REMARKS:

FIG. 3. Score sheet used in the organoleptic evaluation of oils.

Through fortuitous circumstances, the two outside values, one high and the other low, gave balance to the panel. It is interesting to note from the distribution of the standard deviations (Fig. 5) that only one person was erratic in scoring. As a result of this analysis the work of the two tasters who had high and low scores, and of the one who lacked reproducibility are being watched. If further tests confirm their inability to stay within the panel limits, it can be recommended that they be removed from the panel.

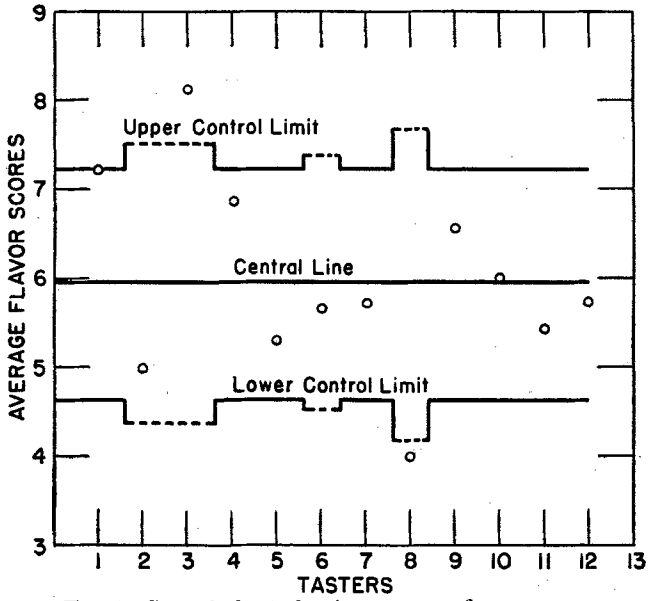

FIG. 4. Control chart showing average flavor scores.

The other method of measuring individual performance is that of comparing the correlation coefficients,[2] (r), and regression coefficients,[2] (b) when the individual's score has been correlated with the average score of the remainder of the panel (7). A large number of samples, ranging from 67 to 106, were used to obtain each coefficient (Table I). The range of

TABLE I.

Correlation of Individual Scores With Average of Remainder of Panel

Taster	Number of samples	Correlation coefficient	Regression coefficient	Standard error of regression
1	88	0.76	0.94	1.3
2	98	0.73	0.94	1.5
3	88	0.79	1.13	1.3
4	67	0.82	0.91	1.0
5	82	0.67	1.07	1.9
6	100	0.76	0.76	1.0
7	72	0.85	1.28	1.2
8	106	0.83	1.15	1.2
9	101	0.78	0.75	0.9
10	94	0.84	1.42	1.3
11	104	0.53	0.54	1.4
12	92	0.71	0.80	1.2

correlation coefficients is from 0.53 to 0.85 while the regression coefficients range from 0.54 to 1.42. A significant fact is that taster No. 6, with the second lowest correlation coefficient, is the one who showed lack of reproducibility on the control charts. Figures 6 and 6a, respectively, show the scatter diagrams for taster No. 8, who has the highest correlation coefficient, and for taster No. 12, who has the lowest correlation coefficient.

[2] See Appendix B for definitions.

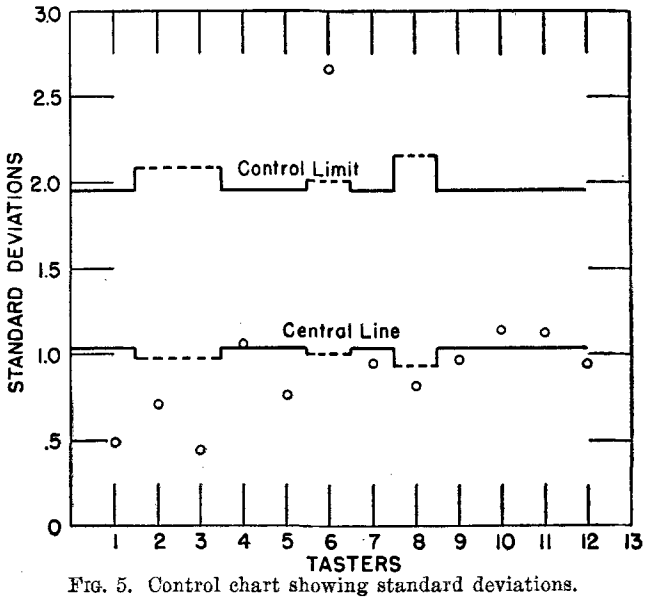

FIG. 5. Control chart showing standard deviations.

Application of these two statistical methods gives measures of a taster's ability and provides an objective basis for dropping certain panel members and substituting others. By selecting tasters on this basis the precision of the taste panel should be improved.

Example of Application. In the course of testing the water-washing, citric-acid process, an experiment was performed which illustrates the applicability of these organoleptic methods. This experiment was designed to answer the question, "What are the effects individually and combined of the water-washing process and the citric-acid-addition process upon flavor stability?" Since two variables were to be evaluated, the design of the experiment called for the refining of four samples. These samples were prepared as indicated in Figure 7. The crude oil and the refining procedures used for these four samples were similar to those previously described (8). Both washed and unwashed crude oils were divided into two parts after completion of the alkali refining, washing, and bleaching steps. To one portion of each of the pairs, citric acid was added during deodorization while to the other portion of each pair no addition was made. The resultant samples are designated as "washed-citrated" (WC), "unwashed-citrated" (UC), "washed" (W), and "unwashed" (U). Six tastings were required at each of the storage periods to evaluate the four oils by the paired-sample technique. Thus each sample was tasted in combination with three other samples at each storage period. The results of these tastings (Table II) are exemplary of the reproducibility of the panel. Sample WC stored 3 days at 60° C. was graded 7.1, 7.1, and 6.8 (av. 7.0) when tasted on 3 different days and when paired with the three other samples. UC was graded 6.6, 6.7, and 6.2 (av. 6.5); W was graded 7.0, 5.8, and 6.1 (av. 6.3); and U was graded 5.3, 4.7, and 4.6 (av. 4.9).

TABLE II.

Flavor Scores and Significance of Differences of Samples Stored 0 and 3 Days at 60° C.

Samples	Stored	
	0 days	3 days
W vs. U	8.5 vs. 7.8+	7.0 vs. 5.3*
WC vs. UC	7.8 vs. 8.1+	7.1 vs. 6.6+
WC vs. W	7.2 vs. 8.4+	7.1 vs. 5.8*
UC vs. U	8.4 vs. 7.8+	6.7 vs. 4.7**
WC vs. U	7.8 vs. 7.9+	6.8 vs. 4.6**
US vs. W	8.3 vs. 8.8+	6.2 vs. 6.1+

+ No significant difference.
* Significant difference (5 per cent level).
** Highly significant difference (1 per cent level).

Substantiating previous experiments, the tests showed a highly significant difference between the washed-citrated and the unwashed after storage for 3 days at 60° C. Moreover, in this experiment, a highly significant difference also was found between the unwashed citrated and the unwashed. Between the washed and unwashed and the washed-citrated and washed, significant differences were found. No significant difference was found between the washed-citrated and the unwashed-citrated, also no significant difference was found between the unwashed-citrated and washed. These relationships and the flavor score averages, calculated from the three determinations, are shown in Table III.

The following conclusions may be drawn from this experiment: citric acid and water washing by themselves improve the flavor stability of edible soybean oil. However, when citric acid has been added to the deodorizer, the increase in stability due to water washing does not give rise to a significant difference.

Caution is advised in the application of statistical techniques to insure correct and valid conclusions. In

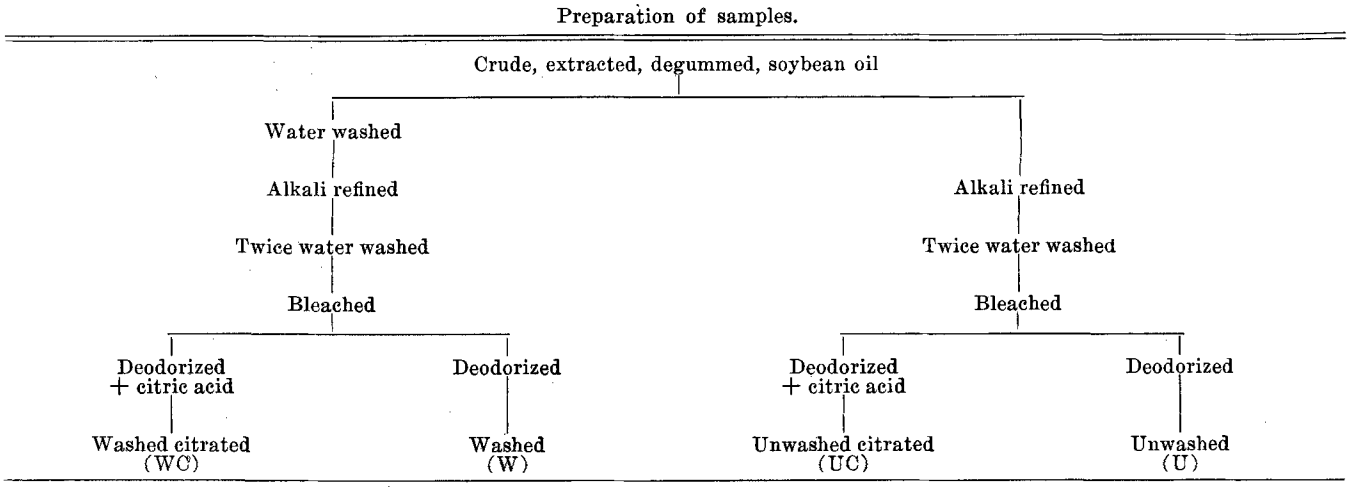

FIG. 7. Diagram showing the preparation of samples.

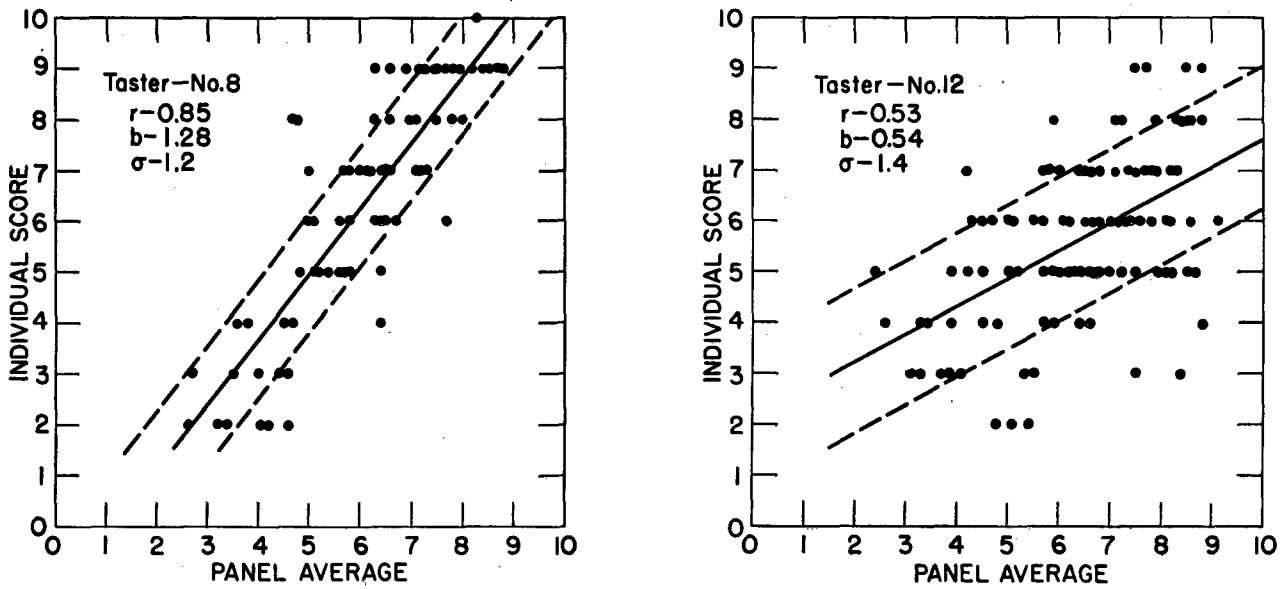

FIG. 6. Scatter diagrams for the taster (No. 8) having the highest correlation coefficient (Fig. 6) and for the taster (No. 12) having the lowest correlation coefficient (Fig. 6a).

the present case, for example, it is assumed that all the oils received the same basic treatment, and that the only differences lay in those refining procedures specifically under test. If some gross change in refining procedure should be made unknowingly, the final results by the panel members might indicate a significant difference that was spurious. Therefore, it is imperative to recognize the need of knowledge of conditions under which the experimentation is done. Repeating the experiment under similar conditions will serve as a check.

A further observation on the application of statistical methods can be made. Although a difference does not reach the arbitrary level of significance generally used in statistical analysis (the 5% level), this does not imply that no difference exists; it merely indicates that the difference, be it real or not, is smaller than the "experimental error," or, in other words, is sufficiently small that it may be accounted for by chance. In fact, in the example just presented, the peroxide values of the samples tasted and the peroxide values of samples held under the conditions of the Swift stability test (9) for 8 hours indicate that the differences which are below the 5% level of significance are none the less real.

It is apparent also from Table III that the same relative order is obtained by the objective chemical tests of stability as by the organoleptic method of evaluation. Because the trends in organoleptic data are generally paralleled by those in peroxide values and because of the inherent consistency and reproducibility of the results of this panel, confidence has been gained in the organoleptic method.

TABLE III.

Summary of Stability Tests.

	WC	UC	W	U
Av. Flavor Score (3 days, 60° C.)	(7.0)	(6.5)	(6.3)	(4.9)
Peroxide Value (3 days, 60° C.)	1.6	1.8	6.6	8.6
Swift Stability Test (P. V. at 8 hours)	9.1	21.0	49.0	86.2

* No significant difference.
* Significant difference (5 per cent level).
** Highly significant difference (1 per cent level).

Acknowledgment

The authors are indebted to Robert Beal of the Engineering and Development Division for his assistance in the refining operations, and to Raymond F. Paschke, formerly of the Oil and Protein Division, for designing the aluminum-topped table and blocks used to warm the samples.

Appendix A: The "t" test

While there are several methods for determining statistical significance, the one applicable to testing the significance of the difference between two means of small samples is the "t" test. If the result of the "t" test is "significant" or "highly significant," it indicates that the difference between the means is too great to be ascribed solely to chance, and that other significant factors are causing the difference.

The "t" test is simple, as there are just 2 formulae needed, which can be calculated rapidly. It is necessary to find the variance (S^2, or square of the standard deviation) of each sample, and then calculate "t".

A convenient formula for the machine calculation of the variance when less than 25 items are involved in an individual sample is:

$$S^2 = \frac{\Sigma X^2}{(N-1)} - \frac{(\Sigma X)^2}{N(N-1)} \qquad (1)$$

The general formula for "t" can be written:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\left[\left(\frac{N_1 + N_2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 N_2}\right)\right]^{1/2}} \qquad (2)$$

with degrees of freedom $= N_1 + N_2 - 2$

When there are the same number of items in the two groups being compared, as in our case of "paired"

samples, ($N_1 = N_2 = N$), the formula may be written:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\left(\frac{S_1^2 + S_2^2}{N - 1}\right)^{1/2}} \qquad (3)$$

with degrees of freedom $= 2(N - 1)$

A glossary of the above symbols follows:

X = individual item

ΣX_1 = sum of the items in the first sample

ΣX_2 = sum of the items in the second sample

ΣX^2 = sum of the squares of the individual items

$(\Sigma X)^2$ = square of the sum of the items in a sample

N_1 = number of items in first sample

N_2 = number of items in second sample

$\overline{X_1} = \frac{\Sigma X_1}{N_1}$ = mean of the first sample

$\overline{X_2} = \frac{\Sigma X_2}{N_2}$ = mean of the second sample

S^2 = variance of the sample, or the square of the standard deviation of the sample.

An example of testing the significance of the difference between two means is given in detail:

Taster No.	Oil 1 (X_1)	Oil 2 (X_2)
1	9	10
2	9	9
3	9	10
4	9	9
5	6	9
6	9	9
7	7	9
8	9	9
9	10	10
10	8	8
ΣX	85	92
$\frac{\Sigma X}{N}$	8.5	9.2
ΣX^2	735	850
$(\Sigma X)^2$	7225	8464

Using (1)

$$S_1^2 = \frac{735}{9} - \frac{7225}{90} = 1.3889$$

$$S_2^2 = \frac{850}{9} - \frac{8464}{9} = .4000$$

and by (3)

$$t = \frac{8.5 - 9.2}{\left(\frac{1.3889 + .4000}{9}\right)^{1/2}\left(\frac{1.7889}{9}\right)^{1/2}} = \frac{-.7}{} = -1.57$$

with degrees of freedom $2(10-1) = 18$.

To interpret this result, it is necessary to refer to a table of "t" values such as Table II of "Statistical Methods for Research Workers" by R. A. Fisher (see No. 10 under References). We find that for $t = -1.57$ with 18 degrees of freedom that P (the probability) $= 0.14$. In other words, there are about 14 chances in 100 of observing a greater difference in mean values. The most usual criteria used in judging statistical significance are: If the probability is greater than 0.05, the differences measured are said to be "not significant." When the probability falls between 0.05 and 0.01, it is considered "significant," and anything less than 0.01 is "highly significant."

Appendix B: General statistical terms

r.—Correlation coefficient. A measure of the degree of variation in one variable which is associated with a given change in another variable. The values range from −1 to 0 to +1, with ±1 indicating perfect correlation, and 0 the absence of any correlation. In this paper, a high correlation coefficient indicates that the individual's score is closely associated with the average score of the rest of the panel. Conversely, a low correlation coefficient indicates that the individual's score does not agree, or is not closely associated with the average of the rest of the panel.

b.—The regression coefficient known also as the slope of the regression line, shows the average rate of change of the one variable corresponding to a unit change in the other variable.

σ.—The standard error of regression is a measure of the scatter of the points about the regression line.

REFERENCES

1. Platt, W. Food Res. *2*, 237 (1937).

2. King, F. B. Food Res. *2*, 207 (1937).

3. Experimental Procedure for Conducting Taste and Smell Tests. U. S. Department of Agriculture, Bureau of Human Nutrition and Home Economics, mimeo, release. (Nov. 1943).

4. Snedecor, G. W., "Statistical Methods," 4th Ed. Collegiate Press, Inc., Ames, Ia. (1946).

5. Guide for Quality Control and Control Chart Method of Analyzing Data. Am. Stand. Assoc. Z1.1, Z1.2, N. Y. (1941).

6. Marcuse, Sophie, "An Application of the Control Chart Method to the Testing and Marketing of Foods." J. Am. Stat. Assoc., Vol. 40, pp. 214-222 (1945).

7. Hopkins, J. W. Canad. Jour. Res. 24F, 203 (1946).

8. Dutton, H. J., Moser, Helen A., and Cowan, J. C. J. Am. Oil Chemists' Soc. *24*, 261 (1947).

9. King, A. E., Roschen, H. L., and Irwin, W. H. J. Am. Oil Chemists' Soc. *10*, 105 (1933).

10. Fisher, R. A. "Statistical Methods for Research Workers," Oliver and Boyd (1925-1944).