STATISTICAL METHODS IN THE COSMETIC INDUSTRY: AN INTRODUCTION

By A. W. MIDDLETON, B.Sc., Ph.D., F.R.I.C.*

THIS ARTICLE is an attempt to give cosmetic workers an outline of a very few of the possible applications of statistical methods to their work. Space does not allow of any very fundamental approach to the subject, but it is hoped that readers will be able to apply the technique of the examples given to similar problems of their own and that they will be encouraged to make use of the bibliography for further study.

* * *

Statistical methods have very wide application and are the means by which the maximum amount of information can be obtained from a given set of experimental results consistent with any desired degree of probability. The general principle of statistical analysis for the judgment of the significance of any difference between two or more sets of observations is the adoption of the "Null Hypothesis"—that there is no difference between the sets of observations—and the subsequent calculation of the probability that in any particular case the variations found between the sets of observations could have been obtained by chance. If this chance is small, the conclusion is drawn that the sets of observations do differ and that the different treatments given the sets have a significant effect. It can be considered that a significant difference between sets of observations exists when it is probable that only once in twenty times would the particular data have been obtained by chance ($p = 0{\cdot}05$). Other commonly used "levels of significance" are probabilities that the results would be obtained by chance once in a hundred times ($p = 0{\cdot}01$) and once in a thousand times ($p = 0{\cdot}001$), the level being chosen by the experimenter to meet the particular degree of accuracy that he requires. For most cosmetic work the writer has found $p = 0{\cdot}05$ satisfactory.

For easy reference, the symbols used and the tabular data required for the proper understanding of the techniques to be described are collected together and shown on the opposite page.

* Chesebrough Manufacturing Co. Ltd., London.

SYMBOLS USED

$x_{1,2,3} \ldots$ $y_{1,2,3} \ldots$	Any individual result or series of results.
p	Probability that an event will occur.
q	Probability that an event will not occur ($q = p - 1$).
σ^2 (Sigma Squared)	Variance ($\sigma^2 =$ sum of squares of deviations from the mean/degrees of freedom).
$d.f.$	Degrees of freedom.
Σ (Sigma)	The sum of a series of results. Thus Σx is the sum of all the results $x_1 + x_2 + x_3 + \ldots$
n, N	The number of results in a series.
$\bar{x}$	The mean of a series of results $x_{1,2,3}, \ldots$
$\bar{y}$	The mean of a series of results $y_{1,2,3}, \ldots$
σ	Estimated standard deviation (square root of the variance).
σ_m	Estimated standard deviation of the mean of a series, often called "Standard Error" ($\sigma_m = \sigma/\sqrt{n}$).
d	Deviation. Difference between a result x and the mean of the series $\bar{x}$.
t	Student's t. The deviation divided by its estimated standard error ($t = d\sqrt{n}/\sigma$).
$S.S.$	Sum of Squares.
$C.T.$	Correction Term.
$M.S.$	Mean Square.
F	The ratio between two variances, the larger estimate of variance being the numerator.
$!$	Factorial (thus $4!$ is factorial $4 = 4 \times 3 \times 2 \times 1$).
χ^2 (Chi Squared)	The sum of squares of a number of variables which vary normally and independently about zero with unit variance.
f_e	Frequency expected.
f_o	Frequency observed.
r	Correlation coefficient.
ρ (rho)	Spearman's ranking coefficient.
W	Coefficient of Concordance.
m	Average frequency.
e	Base of natural logarithms ($e = 2{\cdot}71828$).
$>, <$	Greater than, smaller than (A is greater than B : $A > B$).

Values of t required by the Examples Given

Degrees of Freedom	Probability 0·8	0·7	0·1	0·05	0·01
3	0·277	0·424	2·353	3·182	5·841
5	0·267	0·408	2·015	2·571	4·032
10	0·260	0·397	1·812	2·228	3·169
11			1·796	2·201	3·106
15			1·753	2·131	

Values of F required by the Examples Given

N_1 is degrees of freedom of larger variance.
N_2 is degrees of freedom of smaller variance.

	$p = 0{\cdot}20$		$p = 0{\cdot}001$
N_2 \\ N_1	4	5	2
10	1·8	1·8	14·9
15	1·7	1·7	11·3
18	1·67	1·64	10·4
19	1·66	1·63	10·2

Values of χ^2 required by the Examples Given

Degrees of Freedom	Probability 0·2	0·1	0·02	0·01
1	1·642	2·706	5·412	6·635
2	3·219	4·605	7.824	9·210
4	5·989	7·779	11·668	13·277

Value of r required by Example Given

Degrees of Freedom	Probability 0·01	0·001
2	0·990	0·999

(More complete tables can be found in the works listed at the end of this article.)

Variance

There are several ways in which the variability of a set of results can be expressed, of which the more useful are the "range" and the "variance." The range of a series of results is the difference between the highest and the lowest result of the series.

The variance is the sum of the squares of the deviation of each individual result from the arithmetical mean of the series, divided by the number of results in the series when there are 30 or more results, divided by one less than the number of results when the series consists of less than 30 results.

The variance is universally denoted by σ^2 (sigma squared).

The divisor of the squares of the deviations is called the "degrees of freedom," associated with the variance and can be regarded as the number of deviations from the mean that can be selected at random. Since the sum of the deviations from the mean of n results must be zero, only $n - 1$ of the results can be selected at random, the series having then $n - 1$ degrees of freedom. Symbolically,

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

The standard deviation is the square root of the variance and is thus denoted by σ.

The uses of the standard deviation and of the variance will become evident from a study of the examples of the application of statistical methods which follow.

Calculation of Variance

Consider a series of results obtained for the weight of the contents of bottles being filled on an automatic filling line. These were 37·6, 38·2, 35·8, 36·2, 36·4, 37·3, 37·4, 36·9, 35·8, 36·7, 36·7, 36·3 grams. It is required to know the mean bottle content of the whole of the population (i.e., the whole of the bottles filled under the conditions of sampling), from which the samples weighed were selected at random during the filling, and further, the limits of the mean such that the probability is 19 times out of 20 times that the true mean of the population will be within these limits.

By definition the variance of the sample

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

The calculations are best tabulated, giving :

	x	$x - \bar{x}$	$(x - \bar{x})^2$
	37·6	0·825	0·680625
	38·2	1·425	2·030625
	35·8	− 0·975	0·950625
	36·2	− 0·575	0·330625
	36·4	− 0·375	0·140625
	37·3	0·525	0·275625
	37·4	0·625	0·390625
	36·9	0·125	0·015625
	35·8	− 0·975	0·950625
	36·7	− 0·075	0·005625
	36·7	− 0·075	0·005625
	36·3	− 0·475	0·225625
Sum (Σ) =	441·3	0·000	6·002500

No. of Results (n) = 12.
Mean ($\bar{x}$) = 36·775.

$$\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{6{\cdot}0025}{12 - 1} = \frac{6{\cdot}0025}{11} = 0{\cdot}54568$$

Note that $x - \bar{x}$ is checked by addition, when it should total zero.

It is seen that the arithmetic is laborious even with the aid of a table of squares. However, this work can be reduced considerably by "coding" the results by deducting a convenient value and multiplying to remove the decimal point. The mean of the coded values is calculated and transformed into the original units by reversing the coding process to give the true mean. The variance is calculated by making use of the algebraically equivalent formula for it, i.e.,

$$\sigma^2 = \frac{\Sigma(x)^2 - (\Sigma x)^2/n}{n - 1}$$

The calculations are then :

	x	Coded results, i.e. $(x - 35{\cdot}0)$ 10 Called Y	Y^2
	37·6	26	676
	38·2	32	1,024
	35·8	8	64
	36·2	12	144
	36·4	14	196
	37·3	23	529
	37·4	24	576
	36·9	19	361
	35·8	8	64
	36·7	17	289
	36·7	17	289
	36·3	13	169
$\Sigma =$	441·3	213	4,381

$$\sigma^2_{\text{coded}} = \frac{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}{n - 1}$$

$$= \frac{4{,}381 - \frac{213^2}{12}}{11}$$

$$= \frac{4{,}381 - \frac{45{,}369}{12}}{11}$$

$$= \frac{4{,}381 - 3780{\cdot}75}{11}$$

$$= \frac{600{\cdot}25}{11} = 54{\cdot}568$$

When the coding involves a multiplication factor (in this case 10), the variance must be divided by the square of that factor to give the variance of the original data, thus :

$$\sigma^2 = \frac{54{\cdot}568}{100} = 0{\cdot}54568$$

in agreement with the value found directly. The mean of the coded results is :

$$\frac{213}{12} = 17{\cdot}75$$

which must be divided by the multiplier (10) and 35·0 added to the result to give the mean of the original data, i.e.,

$$\frac{17{\cdot}75}{10} + 35{\cdot}0 = 36{\cdot}775$$

The Standard Deviation is the square root of the variance, in the above example being

$$\sqrt{0{\cdot}54568} = 0{\cdot}7386, \text{ or } 0{\cdot}74$$

for all useful purposes. This relates to any individual result. The standard deviation of the mean of the series of results is given by the formula

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

$$\text{Thus our } \sigma_m = \frac{0{\cdot}74}{\sqrt{12}} = 0{\cdot}21$$

If the distribution of the results is such that there are a maximum number about the mean, tailing off on either side of it, then in practice the distribution can be assumed to be normal (or Gaussian), for which tables have been constructed showing how often a variation (d) from the mean expressed in terms of the standard deviation of the mean can be expected to occur as a matter of chance. This relation d/σ is termed Student's t, and tables of this value are quoted in most standard works on statistics.

In this example, it is desired to know the limits of the true mean of the population at a probability level $p = 0{\cdot}05$.

Reference to a table of Student's t (see above) shows that for the example with 11 degrees of freedom, at $p = 0{\cdot}05$, t is 2·20.

$$\text{Now, by definition, } t = \frac{d}{\sigma_m}, \text{ thus } 2{\cdot}20 = \frac{d}{0{\cdot}21}$$

$\therefore d = 2{\cdot}20 \times 0{\cdot}21 = 0{\cdot}462$, which is the deviation either side of the mean within which the probability is $\frac{19}{20}$ that the true mean of the population lies.

Our mean is 36·775 and the limits are therefore 36·775 ± 0·462: thus the true mean at $p = 0{\cdot}05$ lies between 36·313 and 37·237 grams.

Similarly the limits of $p = 0{\cdot}001$ are between 35·843 and 37·707 grams. These limits of the population mean are called the "fiducial limits" at whatever probability they are calculated: hence, the fiducial limits at $p = 0{\cdot}05$ are 36·313/37·237 grams, and once in 20 times a sample of 12 drawn at random will give a mean which will not lie within these fiducial limits.

Use of "*t*"—Comparison of Results

(a) Results which can be converted to a single series of differences.

Two series of results can be compared with each other very easily by using the *t* function. For example, suppose there to be six samples of an essential oil, each of whose alcohol content has been determined by two different methods. It has to be decided whether there is sufficient evidence to establish that a true difference exists between the two methods at $p = 0{\cdot}05$ level.

By definition, $t = \dfrac{\text{mean of a series of observations}}{\text{standard deviation of the mean (Standard Error)}}$; thus its calculation amounts to the calculation of the mean difference between the samples due to the method of analysis and of its standard deviation. The analytical results and the details of the calculation follow.

Sample No.	Method 1 % Alcohol	Method 2 % Alcohol	Difference, x (Col. 2 − Col. 3)	x^2
1	26	15	11	121
2	8	10	− 2	4
3	14	6	8	64
4	24	31	− 7	49
5	8	19	− 11	121
6	17	22	− 5	25
		Sum	− 6	384

N = 6

$$\text{The mean difference} = \frac{-6}{6} = -1$$

$$\sigma^2 = \frac{\Sigma(x^2) - (\Sigma x)^2/N}{N-1} = \frac{384 - (-6)^2/6}{5} = \frac{384 - 6}{5} = \frac{378}{5} = 75{\cdot}6,$$

and the standard deviation of the mean, $\sigma_m = \dfrac{\sigma}{\sqrt{N}} = \dfrac{\sqrt{75{\cdot}6}}{\sqrt{6}} = \sqrt{12{\cdot}3} = 3{\cdot}51.$

Making the hypothesis a "null hypothesis" that there is no difference between the two methods, there has to be found the probability that if this was so the series of results obtained were obtained by chance. (If the results would not be likely to be due to chance—say, $p = 0.05$ or lower—then there must be a difference due to the method.)

$$t = \frac{d}{\sigma_m} = \frac{-1}{3{\cdot}51} = -0{\cdot}29.$$

Disregarding the sign of t and referring to the table of its values with 5 *d.f.* shows that at $p = 0{\cdot}8$, $t = 0{\cdot}27$. Thus, the probability that if the methods were equal this difference would occur by chance is $p = 0{\cdot}8$, or four in five, and it is very probable that there is no real difference between the methods.

The same samples were analysed by a third method to give the results :

Sample No.	Method 1 % alcohol	Method 3 % alcohol	Difference (1 − 2) x	x^2
1	26	52	− 26	676
2	8	24	− 16	256
3	14	35	− 21	441
4	24	43	− 19	361
5	8	34	− 26	676
6	17	48	− 31	961
		Sum	− 139	3,371

$$\text{Mean difference,} = \frac{-139}{6} = -23{\cdot}2.$$

$$\sigma^2 = \frac{3{,}371 - (139)^2/6}{5} = 30{\cdot}2.$$

$$\sigma_m = \sqrt{\frac{30{\cdot}2}{6}} = 2{\cdot}24.$$

$t_{(5\ d.f.)} = \dfrac{23{\cdot}2}{2{\cdot}24} = 10{\cdot}4$ for which p is very much smaller than $0{\cdot}001$ (written $p << 0{\cdot}001$) and it is certain that these two methods give different results. It is possible to estimate the fiducial limits of this difference, e.g., at $p = 0{\cdot}05$, $t_{(5\ d.f.)} = 2{\cdot}57$, and the limits are $-23{\cdot}2 \pm 2{\cdot}57 \times 2{\cdot}24 = 23{\cdot}2 \pm 5{\cdot}75 = 17{\cdot}45$ to $28{\cdot}95$. It can be expected that Method 3 will give results between 17·45 and 28·95 per cent higher than Method 1 (or Method 2, since this has previously been found equal to Method 1).

It will be seen later that there are shorter ways of comparing three or more series of results (see Analysis of Variance (a)).

(b) Results which must be treated as two series.

In the determination of alcohol just considered, analyses were carried out on the same sample of oil by two methods, and the difference between the results is a measure of the difference due to the variation between methods. However, if the results had been obtained on 12 different samples, six being analysed by Method 1 and six by Method 2, this difference method could not have been used due to the introduction of the variation between the samples. The results are then analysed using the formula :

Let $x_1, x_2 \ldots\ldots$ be results by Method 1, of which there are N results.

$y_1, y_2 \ldots\ldots$ be results by Method 2, of which there are n results.

$$\text{Then } \sigma^2 = \frac{\left[\Sigma(x^2) + \Sigma(y^2) - \frac{(\Sigma x)^2}{N} - \frac{(\Sigma y)^2}{n}\right]}{N + n - 2}$$

$$\text{and } t = \frac{\bar{x} - \bar{y}}{\sigma} \sqrt{\frac{Nn}{N + n}}.$$

The calculation details are :

	Method 1 x	x^2	Method 2 y	y^2
	26	676	15	225
	8	64	10	100
	14	196	6	36
	24	576	31	961
	8	64	19	361
	17	289	22	484
Σ	97	1,865	103	2,167

$$\bar{x} = \frac{97}{6} = 16{\cdot}2$$

$$\bar{y} = \frac{103}{6} = 17{\cdot}2$$

$$\frac{(\Sigma x)^2}{N} = \frac{97^2}{6} = \frac{9{,}409}{6} = 1{,}668$$

$$\frac{(\Sigma y)^2}{N} = \frac{103^2}{6} = \frac{10{,}609}{6} = 1{,}768$$

$$\sigma^2 = \frac{1{,}865 + 2{,}167 - 1{,}668 - 1{,}768}{6 + 6 - 2} = \frac{596}{10} = 59{\cdot}6$$

$$\sigma = 7{\cdot}72$$

$$t_{(10\ d.f.)} = \frac{16{\cdot}2 - 17{\cdot}2}{7{\cdot}72} \sqrt{\frac{36}{12}} = \frac{\sqrt{3}}{7{\cdot}72} = 0{\cdot}23$$

In this example there are twelve results in two sets of six ; thus there are $6 + 6 - 2$ results which can be chosen at random, i.e. 10 degrees of freedom. A $t_{(10\ d.f.)}$ of 0·23 corresponds to a probability greater than 0·8 and there is no difference between the two methods.

Analysis of Variance

(*a*) The t test is a rapid method of determining the relation between two factors by using the appropriate procedure described above. However, when three or more factors have to be compared it is laborious to analyse their respective relations in pairs, and the work can be considerably reduced by using "Analysis of Variance" technique.

The total variance, σ^2, of a final result which is affected by several factors has the valuable property of being the sum of the variances due to each single factor contributing to the result. This allows the total variance to be analysed so as to give the variance due to each factor.

When discussing the t test, sections (a, b), an example was given in which the alcohol in six samples of an essential oil was determined by three different methods, and the effect of the method evaluated. It will be useful to use this example again to illustrate the calculation of the simplest case of analysis of variance. For the purposes of this example, it will be assumed that six determinations were made by each of three methods on one bulk of oil.

Determination No.	% Alcohol Content		
	Method 1 x	Method 2 y	Method 3 z
1	26	15	52
2	8	10	24
3	14	6	35
4	24	31	43
5	8	19	34
6	17	22	48
Σ	97	103	236

No. results each method $= 6 = n$.
Total No. of results $= 18 = N$.
Grand Total $= 436 = \Sigma x, y, z$.

These are dealt with by the following procedure:

(1) Square the total of all individuals and divide by the total number of individuals $= \frac{(\Sigma x,y,z)^2}{N} = \frac{(436)^2}{18} = 10{,}561$. (This is the Correction Term ($C.T.$) which is subtracted from each sum of squares.)

(2) Square each individual and add
$= \Sigma x^2, y^2, z^2 = 26^2 + 8^2 + \ldots\ldots + 34^2 + 48^2 = 13{,}846.$

(3) Square the total of each method and divide by the number of individuals in it and add

$$= \frac{(\Sigma x)^2}{n_x} + \frac{(\Sigma y)^2}{n_y} + \frac{(\Sigma z)^2}{n_z} = \frac{97^2 + 103^2 + 236^2}{6} = \frac{75{,}714}{6} = 12{,}619.$$

This leads to the following Analysis of Variance table:

Source of Variance	Degrees of Freedom	Sum of Squares	Mean Square col. 3 ÷ col. 2	Components of Variance
Between Methods	No. of Methods − 1 = 2	(3) − C.T. = 2,058	1,029	$n\sigma_m^2 + \sigma_w^2$
Within Methods	17 − 2 = 15	3,285 − 2,058 = 1,227	81·8	σ_w^2
Total	No. of Individuals − 1 = 17	(2) − C.T. = 3,285		

The Between Methods Mean Square may or may not be significantly different from the Within Methods Mean Square, and this is estimated by means of the F test in which the Mean Square for Between Methods is divided by that Within Methods to give the Variance ratio, F. This is 1,029/81·8 = 12·6. Reference to tables of F, with N_1 the degrees of freedom of the larger Mean Square (two) and N_2 the degrees of freedom of the smaller Mean Square (fifteen), shows that at $p = 0{\cdot}001$, F has a value of 11·3. Our value is greater than this, showing the probability that the value found would be obtained as a matter of chance to be less than 0·001, meaning that there is little doubt that there is a significant difference in the variance between methods when compared with the variance within methods.

The internal "Within Methods" component of the Variance ($= \sigma_w^2 = 81{\cdot}8$) is the measure of the error of the determinations, whilst the "Between Methods" component of variance ($n\sigma_m^2 + \sigma_w^2 = 1{,}029$) gives the variance due to the method. For the latter we have :

$$6\sigma_m^2 + 81{\cdot}8 = 1{,}029$$

$\therefore \sigma_m^2 = 158$. The total variance of the sample is the sum of that due to the method, $\sigma^2 = 158$ and that due to error, $\sigma_w^2 = 82$. It is apparent that the variance due to the method makes a bigger contribution to the total variance than does that due to the error. If it is desired to know the true alcohol content of the oil, more improvement would be given by steps taken to reduce the variance between the methods than by reduction of the error inherent in any one method.

(b) In the t test, section (a), the alcohol was determined by each of three methods, on six samples. To make most use of the data, it can be analysed to show the variance due to the Method, to the Sample and to the Error. The results are as shown in table opposite :

giving the calculations :

(1) Correction Term $= \frac{436^2}{36} = 10{,}561$.

(2) Square each individual and add $= 26^2 + 8^2 + \ldots\ldots = 13{,}846$.

(3) Square the total of each Method and divide by No. of results in it

Sample No.	% Alcohol Content			Σ	Mean
	Method 1	Method 2	Method 3		
1	26	15	52	93	31·00
2	8	10	24	42	14·00
3	14	6	35	55	18·33
4	24	31	43	98	32·67
5	8	19	34	61	20·33
6	17	22	48	87	29·00
Σ	97	103	236	436	
Mean	16·17	17·17	39·33		

and add $= (97^2 + 103^2 + 236^2)/6 = 12,619$.

(4) Square the total of each sample and divide by No. of results in it, and add $= (93^2 + 42^2 + \ldots\ldots)/3 = 34,372/3 = 11,457$.

The Analysis of Variance Table is now :

Source of Variance	*d.f.*	*S.S.*	*M.S.*	Components of Variance
Between Methods	$3 - 1 = 2$	(3) − (1) = 2,058	1,029	$6\sigma_m^2 + \sigma_e^2$
Between Samples	$6 - 1 = 5$	(4) − (1) = 896	179	$3\sigma_s^2 + \sigma_e^2$
Error	$17 - 2 - 5 = 10$	3,285 − 2,058 − 896 = 331	33·1	σ_e^2
Total	$18 - 1 = 17$	(2) − (1) = 3,285		

The variance ratios are measured by reference to the error. Thus the significance of the Between Samples Variance is taken from the variance ratio $179/33{\cdot}1 = 5{\cdot}4$ for 5/10 degrees freedom, whence from tables $p = 0{\cdot}01$ and there is a highly significant difference between the samples. Similarly there is found to be a highly significant difference between the methods. (However, suppose that the variance ratio for the between samples had not been significant, then the *S.S.* for these (896) would have been added to those for error (331) and divided by the degrees of freedom between samples (5) added to those for error (10) to give a new Mean Square for error. The Analysis of Variance table would then have been :

Source of Variance	*d.f.*	*S.S.*	*M.S.*	Components of Variance
Between Methods	2	2,058	1,029	$6\sigma_m^2 + \sigma_e^2$
Error	15	1,227	37·2	σ_e^2
Total	17	3,285	–	

(In this case the Variance ratio, 1,029/37·2 for 2/15 degrees of freedom, is used for testing the significance of the Between Methods Mean Squares.)

The calculation of σ_e^2, σ_s^2, σ_m^2 gives :

$$\sigma_e^2 = 33{\cdot}1,\ \sigma_s^2 = 48{\cdot}7,\ \sigma_m^2 = 166.$$

The mean result for each sample is the mean of three results and the standard deviation of those means is the square root of the error variance divided by the number of individuals used in calculating the mean, i.e., $\sqrt{\frac{\sigma_e^2}{3}} = \sqrt{\frac{33{\cdot}1}{3}} = \sqrt{11{\cdot}03} = \pm 3{\cdot}32$ based on an error with 10 degrees of freedom. From the t table at $p = 0{\cdot}05$, $t_{10\ d.f.} = 2{\cdot}23$. Therefore the $p = 0{\cdot}05$ fiducial limits of a sample mean are $\pm 2{\cdot}23 \times 3{\cdot}32 = \pm 7{\cdot}4$. Thus the true mean for the alcohol content of Sample No. 1 lies between $31{\cdot}0 \pm 7{\cdot}4$ per cent at the level $p = 0{\cdot}05$—i.e., 24 to 38 per cent—and for Sample No. 2, $14{\cdot}0 \pm 7{\cdot}4$ per cent at level $p = 0{\cdot}05$—i.e., 7 to 21 per cent. Thus there is a significant difference between Samples 1 and 2. It is found that the samples fall into two groups : 1, 4 and 6 not differing significantly, at $p = 0{\cdot}05$, from each other, but differing significantly from Samples 2, 3 and 5 which form a second group of samples not significantly differing from each other.

There can also be found the variation between the Methods whose means are the result of 6 determinations. The Standard deviation of each mean is $\sqrt{\frac{\sigma_e^2}{6}} = \sqrt{\frac{33{\cdot}1}{6}} = \sqrt{5{\cdot}52} = \pm 2{\cdot}35$, calculated from an error with 10 $d.f.$ Thus, as before, the fiducial limits at $p = 0{\cdot}05$ level, are $\pm 2{\cdot}23 \times 2{\cdot}35 = \pm 5{\cdot}24$. The limits of the means for the Methods are then :

Method 1 $16{\cdot}17 \pm 5{\cdot}24 = 10{\cdot}9 - 21{\cdot}4$

Method 2 $17{\cdot}70 \pm 5{\cdot}24 = 12{\cdot}5 - 22{\cdot}9$

Method 3 $39{\cdot}33 \pm 5{\cdot}24 = 34{\cdot}1 - 44{\cdot}6$

Methods 1 and 2 are thus not significantly different, but Method 3 differs very significantly from them.

The Binomial Distribution

In a game of dice in which n dice are tossed N times, the frequency with which a six, say, is thrown 1, 2, 3, ... n times in the same throw is given by the terms of the binomial expansion, $N(p + q)^n$, whose rth term is $N \frac{n!}{r!\,\overline{n-r}!} p^r q^{n-r}$, where p is the probability of the event (i.e., throwing a six) occurring and q of it not occurring. Thus throwing 5 dice at a time for 36 times gives a frequency of 5 sixes at one throw of $36 \frac{5!}{5!} \left(\frac{1}{6}\right)^5$, i.e., $\frac{36}{6^5} = 0{\cdot}0046$ times—and a frequency of 2 sixes and 3 other numbers at one throw of

$36 \frac{5!}{2!3!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$ i.e., 36. 10. $\frac{5^3}{6^5} = 5{\cdot}79$ times.

The frequency series that results from the binomial expansion is sometimes useful in practice when assessing the value of subjective observations. Dry rouge can be made by two different methods which give products differing slightly in texture and application. In order to find out whether there is any real difference in consumer preference for one sort of rouge, the two sorts can be coded *A* and *B* and examined by a number of observers (*N*) selected at random from a cross-section of possible customers of the product, who record their overall opinion as "*A* is better than *B*," "*A* is slightly better than *B*," "*A* equals *B*," "*A* is slightly worse than *B*," and "*A* is worse than *B*."

Assuming the "Null Hypothesis," that there is no difference between the rouges, then the frequency with which the various replies would be received as a matter of pure chance is given by the terms of the binomial $N(p + q)^4$, where $p = q = 0{\cdot}5$. (The probability that either will be chosen is one-half, since the rouges are assumed equal.)

$$N(p + q)^4 = N(p^4 + 4p^3q + 6p^2q^2 + 4pq^4 + q^5),$$

giving a frequency distribution of $N\left(\frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16}\right)$. For the purposes of the example, it is assumed that 32 people give an opinion, yielding the following results :

Classification	Frequency Found f_o	Frequency Expected f_e
A is better than *B*	3	2
A is slightly better than *B*	6	8
A equals *B*	8	12
A is slightly worse than *B*	10	8
A is worse than *B*	5	2
Total :	32	32

It is seen that the frequency observed is different from that expected, and this might be held to indicate that *A* is in fact slightly worse (preferred less) than *B*. The probability of this can be examined by means of the chi-squared test.

The χ^2 test (Chi Squared test).

In general, when frequencies are to be compared, the χ^2 test is the technique to be used. χ^2 is defined as $\Sigma\left(\frac{(f_o - f_e)^2}{f_e}\right)$ and reference to tables of this function at the appropriate degrees of freedom gives the probability that the difference between the frequencies is to be expected as a matter of chance when no real difference between them exists.

The calculation of χ^2 from the data above proceeds as follows :

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
	3	2	1	1	0·5
	6	8	− 2	4	0·5
	8	12	− 4	16	1·33
	10	8	2	4	0·5
	5	2	3	9	4·5
Σ	32	32	0		7·33

The degrees of freedom equal the number of classes less 1, = 4.

Reference to a table of χ^2 values (see above) shows that the probability of obtaining a frequency such as that observed here as a matter of chance when no real difference exists is slightly greater than $p = 0{\cdot}10$. This is not a very significant probability, and the conclusion is that B might be preferred, but the evidence is inconclusive. In practice, further results would be obtained. Assume that this has been done to give 64 results with the same distribution frequency.

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
	6	4	2	4	1·0
	12	16	− 4	16	1·0
	16	24	− 8	64	2·7
	20	16	4	16	1·0
	10	4	6	36	9·0
Σ	64	64	0		14·7 = χ^2

The table shows that at 4 *d.f.* a χ^2 value of 14·7 has p slightly less than 0·01, which is decidedly significant and shows that B is certainly to be preferred to A.

In this calculation frequencies have been obtained with expectation below 5. All groups with below 5 observations should be classified with their appropriate neighbours so as to eliminate such small frequencies. Rearrangement of the group gives :

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
18	20	− 2	4	0·2
16	24	− 8	64	2·7
30	20	10	100	5·0
			χ^2	7·9

From tables $\chi^2(2\ d.f.) = 7{\cdot}82$ at $p = 0{\cdot}02$. Therefore there is still a significant difference between preference for A and B. The effect of the elimination of small groups has been to reduce the unbalanced contribution that they make to χ^2; over half the value of χ^2 for the first group of 64 results is contributed by the group with expected frequency 4.

It is not necessary to be able to calculate the expected frequency in order to use the χ^2 test. An example will show this:

Deliveries of bottle caps are received from two suppliers to fit a bottle, obtained from one supplier, which is capped automatically after filling. A sample 3 gross of caps is taken on receipt from the supplier A and bottles capped with them in the regular routine, note being kept of the faulty caps. Similarly 5 gross are inspected from supplier B. Results are:

	Faulty	Satisfactory	Total
Supplier *A*	12	420	432
Supplier *B*	30	690	720
Totals ...	42	1,110	1,152

Is there any significant difference between these two deliveries of caps? The Null Hypothesis is that there is no difference between them, and the probability that the results obtained are due to chance has to be ascertained. The expectation of faulty caps in the first cell is

$$\frac{432}{1{,}152} \times \frac{42}{1{,}152} \times 1{,}152 = 15{\cdot}7$$

Similar calculations are made for the other cells, to give an expectation table:

	Faulty	Satisfactory	Total
Supplier *A*	15·7	416·3	432
Supplier *B*	26·3	693·7	720
Totals ...	42	1,110·0	1,152

Accordingly, $\chi^2 = \frac{3{\cdot}7^2}{15{\cdot}7} + \frac{3{\cdot}7^2}{416{\cdot}3} + \frac{3{\cdot}7^2}{26{\cdot}3} + \frac{3{\cdot}7^2}{693{\cdot}7} = 1{\cdot}71.$

This χ^2 value has only one degree of freedom, for given the marginal totals one can fill only one cell arbitrarily; reference to the table of χ^2 gives for $\chi^2_{1\ d.f.} = 1{\cdot}71$, p is between $0{\cdot}2$ and $0{\cdot}1$ (χ^2 for these values of p is $1{\cdot}64$ and $2{\cdot}71$ respectively). Thus there is no evidence of a significant difference between the caps from the two suppliers although expression of the faults as a per cent of those examined gives:

Faults from $A = 2{\cdot}78$ per cent.
Faults from $B = 4{\cdot}17$ per cent.

which at first glance looks as if B's caps are much worse than A's.

Correlation Coefficient and Line of Best Fit

When a relation between two variables is suspected to exist, it is necessary to determine the best representation of this relation. The data are first examined to see if they will fit the equation for a straight line, $Y = aX + b$, and if they do not, then attempts are made to transform them so that the transformed data will fit a straight line.

Should there be an *a priori* reason to expect the relation to have the forms $Y = ab^X$, or $Y = aX^b$, the data can be transformed by taking logarithms to give $\log Y = \log a + X \log b$ and $\log Y = \log a + b \log X$ respectively. Since a and b are constants, these equations are seen to be linear.

The probability that the straight line fitting the data would have been obtained by chance had there been no valid relation between the variables can be assessed by calculating the correlation coefficient, r.

$$r = \frac{\Sigma(X - \bar{X})\ (Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2\ \Sigma(Y - \bar{Y})^2}}$$

This coefficient can have any value between 1 and -1, depending on whether the straight line has a positive or negative slope. A correlation coefficient of zero indicates that there is no relation between the variables, and the farther from zero that it is the greater the significance of the relation. Tables of the value of r corresponding to its degrees of freedom, which are *two* less than the number of pairs of observations, have been published giving the probability of the value found for r being due to chance when there is really no relation between the variables. The degrees of freedom are two less than the number of pairs of observations, because one degree of freedom is used in fitting the data to the straight line and one degree in fitting to the total.

An example will make the procedure clear. The time taken by a batch of cold cream to cool in still air under standard conditions of stirring is known for vessels of different radius, and it is required to find the relation between radius of the vessel and cooling time, if any relation exists in an easily expressed form.

It is to be expected that the time will be related to the radius by a relation $\log T = a \log R + b$, and plotting the results on log/log paper shows that this is approximately true. The results are therefore transformed by converting them to logarithms and the correlation coefficient calculated from the transformed data as follows:

n = No. of pairs of observations = 4.
T = Cooling time in minutes.

R = Radius in inches.

By definition $r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 (y - \bar{y})^2}}$

	y = Log T	x = Log R	y^2	x^2	xy
	3·1004	1·2175	9·6128	1·4823	3·7732
	2·4065	0·8261	5·7914	0·6824	1·9880
	1·6020	0·3979	2·6262	0·1583	0·6374
	1·4983	0·3284	2·2448	0·1078	0·4920
Σ	8·6072	2·7699	20·2752	2·4308	6·8906
Means	2·1518	0·6925			

Now $\Sigma(x - \bar{x})^2 = \Sigma(x)^2 - \frac{(\Sigma x)^2}{n}$

$$\therefore \Sigma(x - \bar{x})^2 = 2{\cdot}4308 - \frac{(2{\cdot}7699)^2}{4} = 2{\cdot}4308 - \frac{7{\cdot}6722}{4} = 0{\cdot}5127$$

Similarly

$$\Sigma(y - \bar{y})^2 = 20{\cdot}2752 - \frac{(8{\cdot}6072)^2}{4} = 20{\cdot}2752 - \frac{74{\cdot}083}{4} = 1{\cdot}7042$$

$$\Sigma(x - \bar{x})(y - \bar{y}) = \Sigma(xy) - \frac{\Sigma(x)\,\Sigma(y)}{n}$$

$$= 6{\cdot}8906 - \frac{8{\cdot}6072 \times 2{\cdot}7699}{4} = 6{\cdot}8906 - \frac{23{\cdot}8405}{4} = 0{\cdot}9301$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2\, \Sigma(y - \bar{y})^2}} = \frac{0{\cdot}9301}{\sqrt{(0{\cdot}5127)\,(1{\cdot}7042)}} = 0{\cdot}996$$

from which it is seen that there is a probability of less than 0·01 that this series of results would be obtained were there no real relation between the variables. Therefore it is very probable that a straight line relationship $\log T = a \log R + b$ exists. The best straight line fitting such data is that for which the sum of the squares of the deviations from the straight line are a minimum. Our basic equation $y = ax + b$ meets this requirement when

$$a = \frac{\Sigma xy - \frac{\Sigma x\,\Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}, \text{ and } b = y - a\bar{x}$$

This gives

$$a = \frac{6{\cdot}8906 - \frac{8{\cdot}6072 + 2{\cdot}7699}{4}}{2{\cdot}4308 - \frac{(2{\cdot}7699)^2}{4}} = 1{\cdot}814$$

$$\text{and } b = 2{\cdot}1518 - 1{\cdot}814 \times 0{\cdot}6925$$
$$= 0{\cdot}8956$$

∴ the equation is

Log $T = 1{\cdot}814$ Log $R + 0{\cdot}8956$ which by reversing the transformation of the original observations becomes $T = 0{\cdot}8956R^{1{\cdot}814}$

Ranking Methods

It is often convenient to classify a selection of samples in numerical order —e.g., a series of perfumes or flavours in order of preference—and it is desirable to know how closely different observers agree about the numerical order in which they have classified a group of samples. This order is called the rank. From the marking by two observers, one can calculate Spearman's Rank Correlation Coefficient (ρ) using the formula

$$\rho = \frac{1 - 6\Sigma d^2}{n^3 - n}$$

where d^2 is the square of the difference between the rankings of the two observers for any one sample. Like the correlation coefficient, the rank correlation coefficient can have any value between + 1 and − 1, a value of − 1 indicating that one observer's ranks are the exact opposite of the other's. The probability that any given value of ρ will be obtained by chance when there is no valid relation between the two observers' results can be obtained by using tables of Student's t for

$$t = \rho\,\frac{n - 2}{1 - \rho^2} \text{ with } n - 2 \text{ degrees of freedom.}$$

A simple example will suffice :

A series of 5 buying samples of Jasmine Absolute have to be assessed for perfume quality. For this purpose dilute solutions are smelled under standard conditions independently by two skilled observers (unaware of each other's results) and are ranked by them in order of quality. Results and method of calculation are :

	Sample No.					
	A	*B*	*C*	*D*	*E*	Sum
Observer *L*	2	1	3	4	5	
Observer *M*	3	2	5	1	4	
Difference, $M - L$, *d*, ...	1	1	2	− 3	− 1	0
d^2, ...	1	1	4	9	1	16

$$n = 5, \quad n^3 = 125$$

$$\rho = 1 - \frac{6 \times 16}{125 - 5} = 1 - \frac{96}{120} = 1 - 0{\cdot}8 = 0{\cdot}2$$

$$t_{(5\cdot 2,\ d.f.)} = 0{\cdot}2 \sqrt{\frac{5-2}{1-0{\cdot}04}} = 0{\cdot}2 \sqrt{\frac{3}{0{\cdot}96}} = 0{\cdot}35$$

Tables of t show that at 3 *d.f.*, $t = 0{\cdot}35$ has $0{\cdot}8 > p > 0{\cdot}7$, showing that any agreement in ranking by the two observers could quite easily arise by chance.

Sometimes the question arises of the value of several observers' opinion regarding a series of samples. Thus, there may be five assessors who have arranged six samples of a new line which have different perfumes in order of preference of perfume. The assessors in this case would be picked so as to be representative of the customers to whom the line had been designed to appeal and preferably be known to have "sound judgment." Their ranking could be as follows :

Observer	Perfume A	B	C	D	E	F	Sum
L	1	3	2	5	6	4	
M	1	2	3	4	5	6	
N	2	5	3	5	4	6	
P	3	4	6	1	1	2	
Q	3	4	2	1	5	6	
Sum	10	18	16	16	21	24	105
Difference from 17·5 ...	7·5	0·5	1·5	1·5	3·5	6·5	
d^2	56·25	0·25	2·25	2·25	12·25	42·25	117·0

Rankings of this nature are treated by first calculating the coefficient of concordance, W, which can have all values between 1 and 0, the former for complete agreement among the observers, the latter for complete randomness.

Assuming the usual hypothesis that no agreement exists between the judges and finding the probability that these results would then be obtained were that in fact the case (the Null Hypothesis), the sum of the rankings by all five observers for any one sample should be

$$\frac{105}{6} = 17{\cdot}5$$

Let d be the difference between the sum of the rankings found and the expected sum 17·5, N = number of observers and n = number of samples : then

$$W = \frac{12\,\Sigma d^2}{N^2(n^3 - n)}$$

The results give :

$$W = \frac{12 \times 117}{5^2(6^3 - 6)} = 0{\cdot}268.$$

The probability that the value of 0·268 would arise by chance has now to be calculated. This must first be corrected for continuity by subtracting 2 from Σd^2 and adding 2 to the expression $\frac{N^2(n^3 - n)}{12}$ in the calculation of W. This gives :

$$W = \frac{117 - 2}{\frac{5^2(6^3 - 6)}{12} + 2} = \frac{12 \times 115}{5{,}274} = 0{\cdot}262$$

The variance ratio F is then calculated using the formula :

$$F = \frac{(N - 1)W}{1 - W} = \frac{(5 - 1)\,0{\cdot}262}{1 - 0{\cdot}262} = 1{\cdot}42$$

The degrees of freedom for the greater estimate are :

$$(n - 1) - \frac{2}{N} = (6 - 1) - \frac{2}{5} = 4{\cdot}6$$

and for the lesser estimate :

$$(N - 1)\left[(n - 1) - \frac{2}{N}\right] = (5 - 1)\,4{\cdot}6 = 18{\cdot}4.$$

The degrees of freedom will not often be whole numbers, and the value of F will have to be interpolated in the tables. The example gives for $F = 1{\cdot}42$ at 4·6/18·4 *d.f.*, p is greater than 0·20.

This probability of obtaining the agreement between observers which is found in the test shows that there is no justification for claiming that the observers agree as to the best sample and it would be advisable to have the samples examined by a second set of observers and to pool their results with the first set. Should a significant difference then arise, the perfumes could justifiably be ranked in order of their scores (e.g., had this test been significant —i.e., $p = 0{\cdot}1$ or smaller—the perfumes would be ranked *A* first, *B*, *C* and *D* equal second, *E* fifth and *F* sixth).

Quality Control

The setting up of a control system for accepting or rejecting deliveries of goods such as lipstick containers, bottles, caps and any other manufactured article consisting of a large number of single similar articles which are either sound or defective can be best understood by following an example.

Lipstick containers are received in 100 gross lots from their manufacturer, and factory use of them in the past has shown that the delivery average percentage defective is 2. It is required to devise a sampling procedure so that deliveries can be accepted or rejected by inspection of a sample on their receipt and thus reduce the disorganisation on the lipstick assembly line due to occasional lots of containers with a high proportion of rejects. Since the containers have an average of 2 per cent defective, one reject will occur per sample of fifty containers on average, but any given sample may contain 0, 1, 2, 3, etc. defective containers. The expectation of any given number of defective containers is given by the appropriate term of the Poisson Series, i.e. :

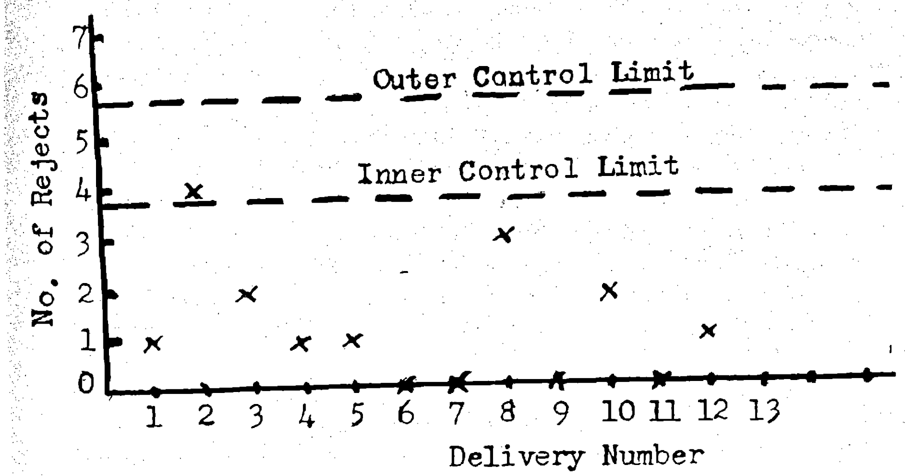
$$e^{-m},\ me^{-m}/1!,\ m^2e^{-m}/2!,\ \ldots\ldots$$

where m is the average frequency. In the case of 50 containers with an average frequency of one defective, the following table shows the expectation calculation and also (1 — cumulative expectation), i.e., the probability that the number of rejects will be more than n.

Samples containing an Average of 1 Defective Each

No. of Rejects n	Poisson Series term	Expectation	Probability that number of rejects will exceed n
0	e^{-1}	0·367879	0·632121
1	$1\ e^{-1}/1!$	0·367879	0·264242
2	$1^2\ e^{-1}/2!$	0·183940	0·080302
3	$1^3\ e^{-1}/3!$	0·061313	0·018989
4	$1^4\ e^{-1}/4!$	0·015328	0·003661
5	$1^5\ e^{-1}/5!$	0·003056	0·000595

The results of the last column show that in a sample of fifty containers selected at random from the 100 gross, 4 or more defective will occur with $p = 0{\cdot}019$ and 6 or more with $p = 0{\cdot}0006$. The numbers 4 and 6 are called the inner and outer limits for defectives respectively, and a graph is drawn on which the rejects occurring in the samples taken from successive deliveries are recorded. This, in the case under discussion, is as follows, the vertical axis giving the number of rejects in a sample of 50 containers :



Should a sample give sufficient rejects to lie outside the outer limits, a further sample of fifty should be examined and the result added to that of the first sample taken from the delivery. The complete sample examined now being made up of one hundred, would normally be expected to contain two rejects (2 per cent of 100). Calculation of the expectation of rejects for this size of sample places the outer limits now at 9 reject containers and the inner limits at 6 rejects. Should 9 or more rejects be found out of the hundred, the delivery is rejected: should 7 or 8 rejects be found, a further sample is taken and the results added to the previous ones and new limits found: should 5 or 6 rejects be found, the delivery would be accepted, but future deliveries watched carefully to see whether a deterioration in the average per cent defective has occurred.

The customary levels of significance for the inner limits of control charts are $p = 0{\cdot}025$ and for the outer limits $p = 0{\cdot}001$. If it is desired to put these levels exactly in the chart (there is no point as a container cannot be less than a unit), then the number of rejects is plotted on a graph against the probability that it will be exceeded, and then the exact value for the number of rejects at $p = 0{\cdot}025$ and $p = 0{\cdot}001$ found by interpolation. In this example these limits are 3·7 and 5·5 defective containers at $p = 0{\cdot}025$, and 5·6 and 8·8 at $p = 0{\cdot}001$, for samples of fifty and one hundred respectively.

Further Study and Applications

No attempt has been made in this paper to do more than show the application of statistical methods to a few typical examples which often occur in the cosmetic industry. Anyone who troubles to study the techniques and their theoretical basis will find that they have acquired additional "laboratory equipment" of similar utility to that of volumetric or chromatographic methods. For practice, readers will find it instructive to re-examine any data from which they have drawn conclusions about which they have not been too confident, by making use of any analogous example given here. They will find the following texts appropriate for further study and not requiring an advanced mathematical background:

"Quality Control Charts," B. P. Dudding and W. J. Jennett, B.S. 600R, 1942.
"Industrial Experimentation," K. A. Brownlee, H.M.S.O., 1949.
"Facts from Figures," M. J. Moroney, a Pelican Book.
"Statistical Methods," G. W. Snedecor, Iowa State College Press.
"Rank Correlation Methods," M. G. Kendall, Charles Griffin & Co., Ltd.
"Statistical Tables for Biological, Agricultural and Medical Research," R. A. Fisher and F. Yates. Oliver & Boyd, Ltd.

A continuation of this article is planned, in which—among other things patch testing and market research techniques will be analysed and the design of experiments dealt with simply.

For easy reference, the examples—together with the statistical procedure in their analysis—are listed below :

For :	*See Section :*
Mean Content of bottles and its fiducial limits.	Standard Deviation. p. 235
Comparison of Methods of analysis :	
Two methods carried out on the same sample.	*t*-test (a). p. 238
More than two methods carried out on one sample, with replication.	Analysis of Variance (a). p. 241
Two methods carried out on different samples.	*t*-test (b). p. 239
More than two methods, each carried out on several samples.	Analysis of Variance (b). p. 242
Subjective assessment of preference for one of a pair of products.	Binomial Distribution, and χ^2 Test. p. 245
Comparison of caps supplied by different manufacturers.	χ^2 Test. p. 247
Relation between cooling time and vessel size.	Correlation Coefficient, and Line of Best Fit. p. 248
Subjective Evaluation of Jasmin Absolute.	Ranking Methods. p. 250
Expert judgment of consumer acceptability.	Ranking Methods. p. 251
Lipstick Containers—control of defectives before factory use.	Quality Control. p. 252

Amended Constitution and Rules

A special general meeting of the Society of Cosmetic Chemists of Great Britain was held on October 23rd, 1952, when general acceptance was obtained for the Society's amended Constitution and Rules. The main difference between the old and new rules lies in the introduction of an associateship for those not fully qualified as cosmetic chemists. The new rules also change the titles of chairman and vice-chairman to president and vice-president respectively—Dr. R. H. Marriott being now the president and Mr. F. V. Wells vice-president.