MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Report S 184

Statistical methods applied to the miximg of solid particles, II

by

R. Doornbos

January
1956

1. Introduction

In a preceding report[1)] a survey was given of statistical problems connected with the mixing of fine-grained powders. Mixtures were considered composed of individual "grains", all having the same size and the same shape, whereas the grains belonging to the different components only differ in colour and possibly in weight.

In this second report a beginning is made with the description of methods which also may be applied when the abovementioned restrictions are dropped. We examine here a statistical method of analysis which is focused on the discovery of a special kind of segregation, namely the gravitational segregation, i.e. a trend of the amount of one component in a vertical direction.

2. Gravitational segregation

We consider a mixture of k components $A_1, \ldots, A_k$ which is divided into n layers by $n-1$ horizontal planes. Now we take from each layer a sample of the mixture which is small in proportion to the total content of the layer. To test whether the component A_1 shows gravitational segregation, we proceed as follows.[2)] In the i^{th} sample we determine the weight- or volume-fraction $\underline{f}_i$ [3)] of A_1-particles. The distribution of $\underline{f}_i$ will depend on the composition of the mixture, on variations in the size of the sample and on the method of analysis used. Because the mixture will not be completely homogeneous in the sense of the definitions a or b of report S 159, the size of the sample expressed in units of weight will vary of the sample is taken e.g. as a level spoonful and finally the determination of the fraction will be subject to errors as the particles cannot be counted in practice but are e.g. submitted to a chemical analysis.

In the following sections methods will be considered for testing the hypothesis H_0, which states that the $\underline{f}_i$ are distributed independently and have the same probability distribution for all $i = 1, \ldots, n$. In practical terms this means, that the k layers are equivalent as to their content of A_1, i.e. that no segregation has taken place.

---

1) J. HEMELRIJK, Statistical methods applied to the mixing of solid particles, I, report S 159 of the Statistical Department of the Mathematical Centre, Amsterdam 1954.

2) This may be done for each of the components separately.

3) Random variables are denoted by underlined symbols.

We especially want to have tests which are powerful against gravitational segregation, i.e. the occurence of large fractions in the lower layers and the contrary of this case, where the large fractions mostly occur in the top layers.

It may be pointed out that the hypothesis H_0 not necessarily implies that the mixing proces from which the mixture has resulted is a random one. If for instance the mixture is clotted as described in report S 159, page 15, where the lumps of the components are randomly mixed, H_0 is still satisfied. If the clotting occurs only in a part of the mixture, e.g. at the bottom of a packet, then the sampling variation of $\underline{f}_i$ will become larger in that place so that H_0 no longer holds. This case will be dealt with in section 8.

3. Coefficients of gravitational segregation

Suppose we have taken a sample in each layer of the mixture and the fractions A_1 found are

$$f_1, \ldots, f_n$$

It is clear that the coefficient

$$g = f_1 + 2f_2 + \cdots + n f_n = \sum_{i=1}^{n} f_n , \tag{3.1}$$

is an adequate measure for segregation in a vertical direction. Because, given n values for the fractions of component A_1 not yet allocated to specific layers, g becomes larger according as the larger values occur in the lower layers.

For practical purposes it is advantageous to use another coefficient g_1 instead of g, g_1 being statistcally speaking equivalent to g, viz.

$$g_1 = \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right) f_i}{\sqrt{\frac{1}{12}(n^3 - n) \sum_{i=1}^{n} (f_i - f_.)^2}} , \tag{3.2}$$

where $f_. = \frac{1}{n}\sum_{i=1}^{n} f_i$. The reasons for adopting g_1 will be explained in the next section.

We get an alternative coefficient if the f_i values are ranked according to their size. Let r_i be the rank of f_i. If some of the f_i are equal the ranks which the equal values would possess if they were different are averaged and this average rank is allocated to each of the tied values. If we replace now in (3.2) f_i by r_i we get another coefficient of gravitational segregation

$$g_2 = \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right) r_i}{\frac{1}{12}(n^3 - n)} . \tag{3.3}$$

The coefficients g_1 and g_2 can only assume values between -1 and +1 and they have the same property as g; they assume larger values if the fractions in the lower layers are large, and low (negative) values if the large fractions occur in high layers.

4. <u>Tests against gravitational segregation, based on g_1</u>

Under the hypothesis H_0 and given n values for $f_1, \ldots, f_n$ found in the samples irrespective of their allocation to the layers, all $n!$ permutations of these values over the layers are equally probable.

Each of these $n!$ allocations gives one value of g_1 and each of these values thus has the probability $\frac{1}{n!}$ to occur. Thus the distribution function of g_1 under H_0, give the values of $f_1, \ldots, f_n$ without their allocation, can be computed exactly. It is immediately clear that the distribution of g_1 is symmetric, for if a permutation of $f_1, \ldots, f_n$ is replaced by the permutation in opposite order, g_1 changes into $-g_1$. Therefore the expected value of g_1, under H_0, is

(4.1) $$\mathcal{E} g_1 = 0 .$$

Further we have 1)

(4.2) $$\begin{cases} \mathcal{E} g_1^2 = \frac{1}{n-1} , \\ \mathcal{E} g_1^3 = 0 , \\ \mathcal{E} g_1^4 = \frac{3}{n^2-1} - \frac{6}{5} \frac{1}{(n+1)(n-1)^2} \left[(n+1) \frac{\sum (f_i - f_.)^4}{\{\sum (f_i - f_.)^2\}^2} - \frac{3(n-1)}{n} \right] . \end{cases}$$

If the last term between square brackets is not too large, a useful approximation of the probability density of g_1 is given by the density:

(4.3) $$\frac{1}{B(\frac{1}{2}, \frac{1}{2}n-1)} (1-g_1^2)^{\frac{1}{2}n-2} , \quad (-1 \leq g_1 \leq 1) .$$

The first four moments of this distribution are 0, $\frac{1}{n-1}$, 0 and $\frac{3}{n^2-1}$ and thus agree closely with those of g_1, if H_0 is true.

---

1) Proofs of the results stated in this section will be given in the appendix.

Further it is known that, under rather weak restrictions g_1 is asymptotically normally distributed (cf. the appendix).

Depending on the value of n one of the following tests may now be applied:

a) $n \leqq 5$.

In this case we use the exact method by computing, if the value of g_1 found is > 0, how many permutations of $f_1, \ldots, f_n$ give a value of g_1 which is as large or larger than the value at hand. If g_1 is negative we want the number of permutations giving values equal to or smaller than g_1. If this number is k, then the two-sided probability of exceedance (two-sided tail probability) of the value of g_1 found is $\frac{2k}{n!}$. I.e.

$$P\left[\,|\underline{g}_1| \geqq g_1 \,\middle|\, H_0; f_1, \ldots, f_n\right] = \frac{2k}{n!},$$

where $\underline{g}_1$ denotes the random variable, under H_0, g_1 the value found in the experiment and $f_1, \ldots, f_n$ the fractions of A_1 in the samples irrespective of their allocation to the layers.

In the following example we have 5 layers in which the fractions of A_1 are

(4.4)	layer	1	2	3	4	5
	fraction	0.17	0.21	0.30	0.29	0.27

The corresponding value of g_1 is 0.793. The permutations giving the largest values of g_1 are

permutation					numerator of g_1	g_1
0.17	0.21	0.27	0.29	0.30	0.34	0.962
0.17	0.21	0.27	0.30	0.29	0.33	0.934
0.17	0.21	0.29	0.27	0.30	0.32	0.906
0.17	0.21	0.30	0.27	0.29	0.30	0.849
0.21	0.17	0.27	0.29	0.30	0.30	0.849
0.17	0.21	0.29	0.30	0.27	0.29	0.821
0.21	0.17	0.27	0.30	0.29	0.29	0.821
0.17	0.27	0.21	0.29	0.30	0.28	0.793
0.21	0.17	0.29	0.27	0.30	0.28	0.793
→ 0.17	0.21	0.30	0.29	0.27	0.28	0.793
0.17	0.27	0.21	0.30	0.29	0.27	0.764

The probability of getting, under H_0, a value as large as or larger than g_1 is therefore $\frac{10}{5!} = \frac{1}{12}$. The two-sided probability of exceedance is thus $2 \times \frac{1}{12} \doteq 0.17$. In this case there is consequently no reason for rejecting H_0.

We may remark here that for the application of this test it is not necessary to compute the g_1 values belonging to the different permutations. It suffices to consider the numerators of g_1, or the values of g as defined by (3.1), which differ from the numerators only by a constant.

b) $5 \leq n \leq 20$.

If n becomes larger than 5, the determination of the exact distribution of $\underline{g}_1$ becomes rather laborious. We can than use the approximation (4.3). When $\underline{g}_1$ has the distribution (4.3), then

$$(4.5) \qquad \underline{t}_1 = \frac{\sqrt{n-2} \cdot \underline{g}_1}{\sqrt{1-\underline{g}_1^2}}$$

has STUDENT's distribution with $(n-2)$ degrees of freedom and we can use the tables of this distribution.

c) $n > 20$.

If n is large we can use the normal (0,1) approximation for the variate $\sqrt{n-1} \cdot \underline{g}_1$.

In the case of our example, where $n=5$, both approximations b) and c) give 0,11 for the bilateral tail probability of $\underline{g}_1 = 0.793$. We see that both approximations give rather poor results in this case, the exact value being 0.17.

5. Tests based on $\underline{g}_2$

The approximation (4.3) for the distribution of $\underline{g}_1$ holds satisfactorily only of the distribution of $\underline{f}_i$ does not depart too much from the normal distribution. If for instance outlying observations occurr among the $\underline{f}_i$ it is better to use $\underline{g}_2$ instead of $\underline{g}_1$. The coefficient $\underline{g}_2$ is identical with the coefficient of rank correlation ρ of SPEARMAN (cf. M.G. KENDALL (1955), p. 21). KENDALL gives a table of the exact distribution of $\underline{g}_2$ up to $n=10$. If $n > 10$ we can use the normal (0,1) approximation for the variate $\sqrt{n-1} \cdot \underline{g}_2$.

Applying this method to our example (4.4) we get $g_2 = 0.60$. Both the exact distribution and the normal approximation give a probability of exceedance (two-sided) of 0.23.

The use of $\underline{g}_2$ has the disadvantage that it is presumably less powerful than that of $\underline{g}_1$, that is a possible gravitational segregation is detected less easily.

6. More than one observation per layer

If we take in the i^{th} layer t_i samples

$$\begin{array}{l} f_{11}, \ldots, f_{1t_1} \\ \vdots \\ f_{n1}, \ldots, f_{nt_n} \end{array}$$

we can apply a test against trend for groups of observations (cf. T.J. TERPSTRA (1952)).[1]

7. The comparisons of two values of $\underline{g}$.

The question also arises to test whether two mixtures have the same gravitational segregation, not necessarily equal to zero. If a number of independent series of observations is available for each mixture, methods are available to test the equality of two (or more) segregations.

Suppose e.g. one is interested in knowing whether shaking of a mixture influences the gravitational segregation. In this case we devide a number of quantities of the mixture at random into two groups and determine the values of $\underline{g}_1$ or $\underline{g}_2$ of both groups after shaking the units of one of the groups during some time.

We then compare the two groups of $\underline{g}$ -values by means of a distributionfree two-sample test, for instance WILCOXON's test (cf. H.B. MANN and D.R. WHITNEY (1947)).

8. Clotting in a part of the mixture

As was pointed out in section 2 the variance of $\underline{f}_i$ will be different ~~fr~~on different values of i if part of the mixture is clotted. Therefore the assumptions used in determining the distributions of $\underline{g}_1$ and $\underline{g}_2$ are then no more satisfied. But if no gravitational segregation has occurred the expected values of the $\underline{f}_i$ will still be the same for all i and therefore

$$(8.1) \qquad \mathcal{E}\underline{g}_1 = \mathcal{E}\underline{g}_2 = 0$$

still holds.

For testing against gravitational segregation we may then analyse a number of packets of the mixture and compute for each packet the value of $\underline{g}_1$. This gives a set of values of $\underline{g}_1$:

$$g_{11}, \ldots, g_{1k}$$

We may then test whether the expectation of the $\underline{g}_{1i}$ $(i=1,\ldots,k)$ is equal to zero by means of STUDENT's test for the mean of a normal distribution. The test statistic of this test is

$$(8.2) \qquad \underline{t} = \frac{\underline{g}_{1.}}{\sqrt{\Sigma(\underline{g}_{1i}-\underline{g}_{1.})^2}} \cdot \sqrt{\frac{k}{k-1}}, \quad (\underline{g}_{1.} = \frac{1}{k}\Sigma \underline{g}_{1i}),$$

----

1) A modified form of this test will be described in a forthcoming paper by the same author. A description (in Dutch) is also given in the memorandum report S 168 (M 61) of the Statistical Department of the Mathematical Centre.

which has, approximately, a Student distribution with $k-1$ degrees of freedom if no segregation is present, assuming approximate normality and equal variances for the variates g_{1i}.

Instead of this test we may also apply a distributionfree test for symmetry (cf. e.g. J. HEMELRIJK (1950)), avoiding the assumptions of normality and equality of variances. This test also has an approximate character, because the distribution of the g_{1i} need not be exactly symmetric.

9. Appendix

Our coefficient g_1 as defined by (3.2) is a special case of the coefficient of correlation r of E.J.G. PITMAN (1937) determined from a set of paired pbservations $x_i, y_1; x_2, y_2; \dots; x_n, y_n$.

$$(9.1) \qquad r = \frac{\sum_i (x_i - x_.)(y_i - y_.)}{\sqrt{\sum_i (x_i - x_.)^2 \sum_i (y_i - y_.)^2}}$$

PITMAN has computed the first four moments of r under the assumption that all $n!$ possible sets of pairs x, y given the numbers $x_1, \dots, x_n$ and $y_1, \dots, y_n$ are equally probable. These moments are

$$(9.2) \quad \begin{cases} \mathcal{E}\, r = 0, \\ \mathcal{E}\, r^2 = \frac{1}{n-1}, \\ \mathcal{E}\, r^3 = \frac{n-2}{n(n-1)} \cdot \frac{h_3}{h_2^{3/2}} \cdot \frac{k_3}{k_2^{3/2}}, \\ \mathcal{E}\, r^4 = \frac{3}{n^2-1} + \frac{(n-2)(n-3)}{n(n+1)(n-1)^3} \cdot \frac{h_4}{h_2^2} \cdot \frac{k_4}{k_2^2}, \end{cases}$$

where

$$h_2 = \frac{1}{n-1} \sum (x_i - x_.)^2,$$

$$h_3 = \frac{n}{(n-2)(n-2)} \sum (x_i - x_.)^3,$$

$$h_4 = \frac{n}{(n-1)(n-2)(n-3)} \left\{ (n+1) \sum (x_i - x_.)^4 - \frac{3}{n}(n-1) \left[\sum (x_i - x_.)^2 \right]^2 \right\}$$

and k_2, k_3 and k_4 are the corresponding expressions in the y_i.

Substituting $x_i = i$ and $y_i = f_i$, we get $r = g_1$ and the moments as given by (4.1) and (4.2)

Further W. HOEFFDING (1952) has shown that under H_0 the test statistic

$$(9.3) \qquad \underline{g}_H = \frac{\sum (a_i - a_.)\underline{f}_i}{\sqrt{\sum (a_i - a_.)^2 \sum (\underline{f}_i - \underline{f}_.)^2}} \cdot \sqrt{(n-1)} \quad ,$$

where $a_1, \ldots, a_n$ are given numbers is asymptotically normally (0,1) distributed if $\mathcal{E} \, |\underline{f}_i|^3 < \infty$ and

$$(9.4) \qquad \lim_{n \to \infty} \frac{\max\limits_{1 \le i \le n} (a_i - a_.)^2}{\sum (a_i - a_.)^2} = 0 \, .$$

If we take $a_i = i$, we get $\underline{g}_H = \sqrt{n-1}\, \underline{g}_1$ and condition (9.4) is satisfied, which proves the asymptotic normality of $\underline{g}_1$, provided $\mathcal{E} \, |\underline{f}_i|^3 < \infty$

Finally E.L. LEHMANN and C. STEIN (1949) have shown that the test based on (9.3) is most powerful for testing th against the alternative that the $\underline{f}_i$ are normally distributed with mean $a_i \xi + \eta$ and common variance σ^2. So our test based on $\underline{g}_1$ is most powerful against the alternative that the $\underline{f}_i$ are normally distributed while the means show a linear trend.

References

J. HEMELRIJK (1950), Symmetrietoetsen, Diss., Den Haag, Excelsior.

W. HOEFFDING (1952), The large sample power of tests based on permutations of the observations, Annals of Mathematical Statistics, 23 (1952), pp 169-192.

M.G. KENDALL (1955), Rank Correlation Methods, second edition, Charles Griffin and Co., London.

E.L. LEHMANN and C. STEIN (1949), On the theory of some non-parametric hypotheses, Annals of Mathematical Statistics, 20 (1949), pp 28-45.

H.B. MANN and D.R. WHITNEY (1947), On a test whether one of two random variables is stochastically larger than the other, Annals of Mathematical Statistics, 18 (1947), pp 50-60.

E.J.G. PITMAN (1937), Significance tests which may be applied to samples from any populations. II The correlation coefficient test, Supplement to the Journal of the Royal Statistical Society, 4 (1937), pp 225-232.

T.J. TERPSTRA (1952), The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, Indagationes Mathematicae, 14 (1952), pp 327-333.
