MULTIPLE COMPARISONS OF OBSERVATION MEANS—
ARE THE MEANS SIGNIFICANTLY DIFFERENT?

T.Z. FAHIDY
University of Waterloo • Waterloo, Ontario, Canada N2L 3G1

Several currently popular methods of ascertaining which treatment (population) means are different, via random samples obtained under each treatment, are briefly described and illustrated by evaluating catalyst performance in a chemical reactor.

In a routine undergraduate classroom application of one-way ANOVA (analysis of variance), the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad (1)$$

is tested against the alternative hypothesis that at least two population means are different, given sample means $\bar{x}_{1.}, \bar{x}_{2.}, \ldots, \bar{x}_{i}$ computed from $n_1, n_2, \ldots, n_k$ random observations. If H_0 is rejected, it is not always *a–priori* obvious which of the two or more means differ; multiple comparison tests help to provide the answer. We are especially interested in the "fate" of H_0 when the individual treatment/population variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ are considered to be equal, because if they are not equal, wide dispersion in the data is most likely the primary cause, and further testing may not be required.

Commonly, the rejection of H_0 is made at a 95%, or at a 99% level of confidence. In the first case, under H_0, the error probability of falsely rejecting H_0 is (at most) 5%, in the second case, it is 1%. This, so-called Type I error is also known as the level of significance: $\alpha = 0.05$ (significant), and $\alpha = 0.01$ (highly significant), respectively. An H_0 maybe rejected at the significant level, but not necessarily at the highly significant level. Consequently, failure to reject H_0 at $\alpha = 0.05$ automatically implies no rejection at $\alpha = 0.01$, as well. The P- value is defined as the smallest level of significance (*i.e.*, the smallest value of α) that would lead to a rejection of H_0. The 5% and 1% levels do not have a privileged position, because the real-life significance of the magnitude of the Type I error is left to the tester's judgment. In the sequel, we shall adhere to the two traditional levels inasmuch as certain statistical tables include only critical values pertaining to $\alpha = 0.05$ and $\alpha = 0.01$ (F – and T – tables are, however, notable exceptions).

MOTIVATION: A CHEMICAL REACTOR WITH DIFFERENT CATALYSTS

In the chemical engineering classroom, we could explore, for instance, the hydrogenation of nitrobenzene[1] in a tubular reactor, using three different Ni-based catalysts. Under identical experimental conditions, conversion data shown in Table 1 (page 19) are assumed to be available for testing H_0 : the three mean conversions are the same. The alternative hypothesis, H_a, is that at least two catalysts render different mean conversions. If H_a is upheld, the next question to ask is which catalysts are significantly different, by testing whether $(\mu_3 - \mu_2)$; $(\mu_1 - \mu_3)$; $(\mu_1 - \mu_2)$ differ significantly from zero. This is the goal of multiple comparison tests, following a standard ANOVA applied to H_0 in Eq. (1).

Three fundamental assumptions are now made: (1) the three catalysts act independently of one another; (2) the observed conversions belong to a normal (Gaussian) population, at least approximately; and (3) the three observation sets are homogeneous, *i.e.*, the conversion-population variances are not significantly different. Since, as demonstrated in Appendix 1, the third assumption is not rejected, we expect that ANOVA will be sufficiently robust, even if the conversion data are only approximately normal.

We shall look at two variations of the theme. The first one involves equal observation sizes $n_1 = n_2 = n_3 = 8$. In the second one we stipulate that four of the observations are absent, *i.e.*, $n_1 = 8$; $n_2 = 6$; $n_3 = 6$. The multiple comparison tests are not the same, as seen in the sequel.

Details of ANOVA can be found in undergraduate statistics textbooks, and are omitted here. Table 1 indicates that rejection of H_0 in Eq. (1) carries only a 0.2%, and a 0.1% Type I error, respectively; we can say that H_0 is soundly defeated. Our inferences about the three catalyst populations (*i.e.*, very large number of observations) are data-specific, of course. With dif-

Thomas Z. Fahidy *is distinguished professor emeritus and adjunct professor of chemical engineering at the University of Waterloo (Canada). He received his B.Sc. and M.Sc. degrees from Queen's (Canada) and his Ph.D. from Illinois (Urbana-Champaign). Recipient of several professional honors including fellowships, he is a Professional Engineer of Ontario, and author of the text* Principles of Electrochemical Reactor Analysis *(Elsevier, 1985).*

ferent sets of observation, the numerical values of the F – statistic and our inferences may well be different.

We now proceed to find out how many of the three catalysts can be expected to behave differently from the others via multiple comparison tests.

MULTIPLE COMPARISONS: EQUAL-SIZE OBSERVATIONS

a. The Bonferroni method[2-4]

This method is based on the concept of the simultaneous level of confidence (SLC), which guarantees that each individual difference in means carries a level of significance α_B as given below:

$$\alpha_B = \frac{\alpha}{k(k-1)} \tag{2}$$

The degrees of freedom for error are $\sum_{i=1}^{k} n_i - k$.

In our reactor, $n_1 + n_2 + n_3 - k = 6$, hence $\alpha_B = 0.0083$ for $\alpha = 0.05$, and $\alpha_B = 0.0017$ for $\alpha = 0.01$. The confidence interval may be written as

$$(x_{i.} - x_{j.}) - t[\alpha_B; k(n-1)]\sqrt{2(MSE)/n} \prec (\mu_i - \mu_j) \prec (x_{i.} - x_{j.}) + t[\alpha_B; k(n-1)]\sqrt{2(MSE)/n} \tag{3}$$

for the difference between two arbitrary population means. The degree of freedom of the T – statistic is k(n-1). We find in T – distribution tables, *e.g.*,[5] the values: t(0.01;21) = 2.518; t(0.005;21) = 2.831; t(0.001;21) = 3.527. By linear extrapolation we obtain t(0.0083;21) = 2.623 and t(0.0017;21) = 3.406. With MSE = 7.786, the confidence intervals in Table 2 are obtained.

What can we conclude from the results? Since $(\mu_3 - \mu_2) = 0$ is included in BCI_{32}, we cannot assert that the two means are significantly different. BCI_{12} and BCI_{13}, however, exclude zero—meaning that μ_1 appears to be significantly different from μ_2 and μ_3. This conclusion is supported even at the highly significant level. In other words, the *mean* performance of Catalyst 1 appears to be different from (in fact, better than) the performance of the other two catalysts, but the *mean* performance of Catalyst 2 does not appear to be significantly different from the performance of Catalyst 3.

b. The Tukey method[6,7]

In this method the statistic Q belongs to the *studentized range distribution*, where the degrees of freedom are $DF_1 = k$ and $DF_2 = k(n-1)$. Critical values of Q[α; k, k(n-1)] are tabulated at $\alpha = 0.05$ and $\alpha = 0.01$,[8-10] but the Q-tables are not widely available in the textbook literature. We need linear interpolation based on the tabulated values q[0.05;3,20] = 3.58; q[0.05;3,24] = 3.53; q[0.01;3,20] = 4.64; q[0.01;3,24] = 4.55, yielding q[0.05;3,21] = 3.57 and q[0.01;3,21] = 4.62. The confidence intervals given by the expression

$$(x_{i.} - x_{j.}) - Q[\alpha; k, k(n-1)]\sqrt{MST/n} \prec (\mu_i - \mu_j) \prec (x_{i.} - x_{j.}) + Q[\alpha; k, k(n-1)]\sqrt{MST/n} \tag{4}$$

are given in Table 2. Agreement with the Bonferroni method is not the same for $(\mu_1 - \mu_3)$, inasmuch as the null hypothesis of no difference between them can be rejected only at the significant, but not at the highly significant, level (if the lower bound of the interval is set to zero, linear interpolation between the two critical values would indicate a P – value of about 0.023.

c. The Scheffé method[11,12]

In establishing the confidence intervals, the Snedecor-Fischer F – distribution provides values of the critical statistic, when two individual means are compared:

$$(x_{i.} - x_{j.}) - \Phi(\alpha; k, n) \prec (\mu_i - \mu_j) \prec (x_{i.} - x_{j.}) + \Phi(\alpha; k, n)$$
$$\Phi(\alpha; k, n) \equiv \sqrt{2(k-1)(MSE)F[\alpha; k-1, n-k]/n} \tag{5}$$

and, specifically in our case,

$$(x_{i.} - x_{j.}) - 0.5\sqrt{(MSE)F[\alpha; 2, 21]} \prec (\mu_i - \mu_j) \prec (x_{i.} - x_{j.}) + 0.5\sqrt{(MSE)F[\alpha; 2, 21]} \tag{6}$$

The significant and highly significant critical F – values are f[0.05;2,21 = 3.47, and f [0.01;2,21] = 5.78 .

The confidence intervals in Table 2 indicate complete qualitative agreement with the Bonferroni method, and nearly complete agreement with the Tukey method.

d. The Duncan method[13,14]

In this approach, differences in the sample means are compared to a statistical parameter to ascertain significance; there are no confidence intervals to deal with. Only equal-size observations can be considered for the test. We have seen that in our case $x_{1.} > x_{3.} > x_{2.}$, hence we have two single ranges (p = 2) and one double range (p = 3) of sample mean differences. If any difference exceeds the numerical value of the Duncan parameter

$$R_p = r_p\sqrt{(MSE)/n} = 0.9865r_p; p = 2,3 \qquad (7)$$

then that difference is at least significant, or even highly significant. The quantity r_p, called the *least significant studentized range*, depends on α and the degree of freedom of MSE = 7.786, *i.e.*, DF = 21. Critical tables of r_p are provided in various textbooks, *e.g.*,[15-17] for $\alpha = 0.05$ and $\alpha = 0.01$. In our case, $r_2 = 2.950$ ($\alpha = 0.05$) and 4.024 ($\alpha = 0.01$), and $r_3 = 3.097$ ($\alpha = 0.05$) and 4.197 ($\alpha = 0.01$) at DF = 20. The next set of critical values given for DF = 24 differing but slightly from the DF = 20 values, interpolation may be bypassed and we accept $R_2 = 2.910$ and $R_3 = 3.055$ at $\alpha = 0.05$, and $R_2 = 3.970$; $R_3 = 4.140$ at $\alpha = 0.01$. The conclusion drawn from this test is that $(\mu_1 - \mu_3)$ and $(\mu_1 - \mu_2)$ are different from zero even at a highly significant level, but H_0: $(\mu_3 - \mu_2) = 0$ is acceptable at a significant (and, of course, at a highly significant) level.

MULTIPLE COMPARISONS: UNEQUAL-SIZE OBSERVATIONS

We now proceed to postulate that the last two conversion measurements with Catalysts 2 and 3, *i.e.*, the entries 81 and 83%, and 80 and 84%, respectively, are unknown. In this instance, the mean conversions are $\bar{x}_{1.}$=84%; $\bar{x}_{2.}$=77%; $\bar{x}_{3.}$=79%, and on account of the computed F – statistic of 15.45, H_0 in Eq. (1) is rejected very strongly, with a Type I error less than 0.1%. We are now ready to investigate the differences in conversion means.

For the establishment of the confidence intervals, there are essentially two modifications with respect to the equal-size observations case. The MSE multiplier for the confidence intervals becomes $\left(n_1^{-1} + n_2^{-1} + \ldots + n_k^{-1}\right)$, and the second degree of freedom is smaller due to the smaller number of observations. In our case MSE = 5.988, and the MSE-multipliers are 7/24 and 1/3. For the Bonferroni method, the critical T – statistics are t[0.00833;17] = 2.678 ($\alpha = 0.05$) and t[0.00167;17] = 3.521 ($\alpha = 0.01$). In the Scheffé method, f[0.05;2;17] =

TABLE 1
Exit conversions in a tubular reactor using three different catalysts. The observations are hypothetical under identical experimental conditions.

Catalyst 1	Catalyst 2	Catalyst 3
82	74	79
86	82	79
79	78	77
83	75	78
85	76	82
84	77	79
86	81	80
87	83	84
Means: 84	78.25	79.75

ANOVA (equal – size)

Source	SS	DF	Mean SS	F
Treatments	142.34	2	71.17	9.14
Error	163.50	21	7.786	
Total	305.84	23		

Critical values of the F – statistic: 3.47 ($\alpha = 0.05$);5.78 ($\alpha = 0.01$); 9.77 ($\alpha = 0.001$);
P – value: ≈ 0.002

ANOVA (unequal – size)
(Last two entries in column 2 and 3 above are removed)

Source	SS	DF	Mean SS	F
Treatments	185.0	2	92.5	15.45
Error	101.80	17	5.988	
Total	286.80	19		

Critical values of the F – statistic: 3.59 ($\alpha = 0.05$); 6.11 ($\alpha = 0.01$);10.66 (α =0.001)
P – value: < 0.001

TABLE 2
Lower (LB) and upper (UB) bounds of the confidence intervals for mean conversions with equal–size observations
B: Bonferroni; T: Tukey; S: Scheffe´

		$\alpha = 0.05$			$\alpha = 0.01$		
Difference	Bound	B	T	S	B	T	S
$\mu_1 - \mu_2$	LB	2.091	2.228	4.401	0.998	1.192	3.646
	UB	9.409	9.272	9.599	10.502	10.308	10.354
$\mu_1 - \mu_3$	LB	0.591	0.728	2.401	0.502	- 0.308	1.646
	UB	7.909	7.772	7.599	9.002	8.808	8.354
$\mu_3 - \mu_2$	LB	- 2.159	- 2.022	- 0.599	- 3.252	- 3.058	- 1.354
	UB	5.159	5.022	4.599	6.252	6.058	5.354

3.59, and f[0.01;2;17] = 6.11. The corresponding confidence intervals in Table 2 indicate complete qualitative agreement among the three methods.

DISCUSSION

What are the relative merits of the three methods? In accordance with pertinent literature, the following statements can be made. (A) The Bonferroni method is "very effective" for small-size comparisons, and if the observation-set sizes are equal, Tukey's method is "optimal" in the sense of yielding the shortest confidence interval.[18] (B) If the product of the two degrees of freedom of the test statistics is large, the smaller of the Bonferroni and Scheffé intervals are to be taken.[19] Tables 1 and 2 are in compliance with (A), excepting the Tukey interval for $(\mu_1 - \mu_3)$ at $\alpha = 0.01$, but this result is not surprising in view of the usually higher power of the Bonferroni method for small k.

One advantage of dealing with confidence intervals with only positive bounds resides in the rejection of the null hypothesis that a difference in the related two population means is not significant. The shorter the confidence interval, the less likely that the lower bound will be negative, therefore H_0: $(\mu_i - \mu_j) = 0$ will be rejected. This is the desired outcome, since rejection of a null hypothesis is a statistically stronger result than failure to reject it.

It follows from what we have already stated at the outset that we would be hesitant about preferring Catalyst 1 if the population variances were shown to be unequal (the case of heteroscedasticity), especially at a very high degree of confidence. Data transformation techniques, *e.g.*,[20] for the removal of heteroscedasticity are known, but they are beyond the objective of this paper.

Multiple comparison methods can readily be extended to contrasts, where arbitrary linear combinations of population means can be treated. A short description is given in Appendix 2.

Finally, we may be interested in comparing our mean conversions with the mean conversion obtained with, for instance, a Ni-free metal oxide catalyst. This exercise would fall into the "treatment comparison with control" test category. In so doing, the hypotheses[21]

$$H_0 : \mu_0 = \mu_i ; i = 1, \ldots, k$$
$$H_1 : \mu_0 \neq \mu_i ; i = 1, \ldots, k \qquad (8)$$

are tested by the Dunnett parameter[22,23]

$$D_1 = \frac{x_{1.} - x_{0.}}{\sqrt{2(\mathrm{MSE})/n}} \qquad (9)$$

where the 0 subscript denotes control. If the absolute value of d_i exceeds the critical value $d[(\alpha/2);k,DF_2]$ with first degree of freedom k, and DF_2, the second degree of freedom of MSE, which includes the control observations, then H_0 in Eq. (8) is rejected at a chosen α. Critical Dunnett parameter values are tabulated for two sided tests, *e.g.*, in Reference 24.

For our reactor, if we assume a control conversion set of 73,80,76,82,77,75,79,83%, we obtain SSE = 247.88 and MSE = 247.88/[4(8-1)] = 8.8528. The values $d_1 = 3.949$; $d_2 = 0.084$; $d_3 = 1.0925$ are computed from Eq. (9). Since at $\alpha = 0.05$, d[0.025;3,24] = 2.51 and d[0.025;3,30] = 2.47, and at $\alpha = 0.01$, d[0.005;3,24] = 3.22, and d[0.005;3,30] = 3.15, only the mean of Catalyst 1 appears to be significantly different from μ_0.

CLASSROOM EXPERIENCE

Respecting the necessity of numerous other topics that also have to be covered in the introductory probability and statistics course taught to second year undergraduate ChE students, this author could go beyond Duncan's multiple-range method only to a rather limited extent in dealing with the subject matter. Class reaction was (not unexpectedly) mixed, depending on the degree of willingness to accept the statistical way of reasoning. Students exposed to statistical techniques during their work terms (the co-op structure of engineering programs at Waterloo alternates in-house lecture terms and practice-based work terms) generally showed more appreciation, if not enthusiasm, than their fellow classmates with different work-term orientation. The author's quest for a course devoted solely to the analysis of variance including multiple comparisons is motivated not only by personal experience arising from the introductory course, but also by the steadily increasing importance of probability-based thinking in all walks of life.

TABLE 3

Lower (LB) and upper (UB) bounds of the confidence intervals for mean conversions with unequal – size observations.

B: Bonferroni; T[&]: Tukey; S: Scheffe´

		$\alpha = 0.05$			$\alpha = 0.01$		
Difference	Bound	B	T	S	B	T	S
$\mu_1 - \mu_2$	LB	3.461	3.607	3.459	2.346	2.570	2.380
	UB	10.539	10.393	10.541	11.654	11.430	11.619
$\mu_1 - \mu_3$	LB	1.461	1.607	1.217	0.346	0.570	0.064
	UB	8.539	8.393	8.783	9.654	9.430	9.936
$\mu_3 - \mu_2$	LB	- 1.783	- 1.627	- 1.541	- 2.975	- 2.736	- 2.619
	UB	5.783	5.627	5.541	6.975	6.736	6.619

[&]modified for unequal-size observation sets [Devore, J.,L., loc.cit., (2004), p.434.]

FINAL REMARKS

Because the statistically informed chemical engineer is especially valuable to industry, education of our students in statistical techniques is highly desirable in today's world. Multiple comparisons make up an integral part of this education in demonstrating the utility of statistical approaches, and the importance of applying proper judgment to test results.

ACKNOWLEDGMENTS

This paper was prepared using facilities provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Waterloo. The unequal-size data in Table 1 were adopted from Hogg and Ledolter[25,26] on ANOVA of beam reflection; the material does not contain multiple-comparison tests.

NOMENCLATURE

B_1;B_2 Bartlett-parameters in Equations (A.4;A.7)
BCI_{ij} Bonferroni confidence interval for $(\mu_i - \mu_j)$
C Bartlett-parameter in Equation (A.6)
c_i contrast coefficient [Eq. (11)]
DF degree of freedom
d two-sided Dunnett's critical parameter [Eq. (13)], at level of significance α
F critical parameter of the Fischer-Snedecor F-distribution, at level of significance α
f numerical value of an F-statistic
G Cochran statistic in Eq. (A.2)
H_a hypothesis alternative to null hypothesis H_0
H_0 Null hypothesis
k number of treatments (observation sets)
LB lower bound of a confidence interval
M Bartlett-parameter in Eq. (A.5)
MSE mean square of experimental errors
MS mean sum of squares
N total number of observations
n number of observations in each equally sized treatment
n_i number of observations in the i – th treatment
Q Tukey – statistic in Eq. (4)
q numerical value of a Q - statistic
r_p least significant studentized range in Duncan's test in Eq. (5)
R_p least significant range of means in Duncan's test in Eq. (5)
S^2 sample variance
SCI Scheffé confidence interval
SS sum of squares
T Student's T – statistic
t numerical value of a T - statistic
UB upper bound of a confidence interval
$x_{i.}$ mean value of x_i observations in the i-th treatment

GREEK SYMBOLS

α level of significance (Type I error, i.e. the maximum value of the probability of rejecting H_0 when H_0 is true)
α_B simultaneous level of significance (Bonferroni method)
μ_i the (true) population mean of observations due to the i – th treatment

REFERENCES

1. Smith, J.M., *Chemical Engineering Kinetics*, 3rd Ed., Example 13-5, p.574, McGraw Hill, NY, (1981)
2. Devore, J.L., *Probability and Statistics for Engineers and the Sciences*, 1st Ed., Section 10.2, p. 357, Brooks/Cole, Monterey (1982)
3. Petruccelli, J.D., B. Nandram, and M. Chen,, *Applied Statistics for Engineers and Scientists*, Section 9.1, p.536, Prentice – Hall, Upper Saddle River, NJ, (1999)
4. Arnold, S.F., *Mathematical Statistics*, Section 13.4, p.477, Prentice Hall, Englewood Cliffs, NJ (1990).
5. Lindley, D.V., and W. F. Scott, *New Cambridge Statistical Tables*, 2nd Ed., Table 10, p. 45, Cambridge Univ. Press (1984)
6. Devore, J.L., and R. Peck, *Statistics—The Exploration and Analysis of Data*, 3rd Ed., Section 13.2, p. 527, Duxbury Press, Pacific Grove CA, (1997)
7. Devore, J.L., *Probability and Statistics for Engineering and the Sciences*, 6th Ed., Section 10.2, p. 422, Brooks/Cole – Thomson Learning, Belmont, CA, (2004)
8. Devore, J.L., and R. Peck, loc. cit., Table VIII, p. 616
9. Devore, J.L., loc. cit. (2004), Table A.10, p. 754.
10. Kokoska, S., and Ch. Nevison, *Statistical Tables and Formulae*, Table 12, pp. 64-65, Springer – Verlag, NY, (1989)
11. Steel, R.G.D., and J.H. Torrie, *Principles and Procedures of Statistics-A Biometrical Approach*, 2nd Ed., Sections 9.2 and 9.3, pp. 196;201, McGraw Hill, NY, (1980)
12. Devore, J.L., loc. cit., (1982), p.366
13. Miller, I.R., J.E. Freund, and R. Johnson, *Probability and Statistics for Engineers*, 4th Ed., Section 12.4, p. 405, Prentice – Hall, Englewood Cliffs, NJ, (1990)
14. Walpole, R.E., R.H. Myers, S.L. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists*, 7th Ed., Section 13.6, pp. 480-481, Prentice Hall, Upper Saddle River, NJ, (2002)
15. Miller, I.R., et al., loc. cit., Table 12(a), p. 585, Table 12(b), p. 586
16. Walpole, R.E., et al., loc. cit., Table A.12, pp. 688-689
17. Dougherty, E. R., *Probability and Statistics for the Engineering, Computing and Physical Sciences*, Table A.9, pp. 741-742, Prentice Hall, Englewood Cliffs, NJ, (1990)
18. Petruccelli, J.D., et al., loc. cit., p. 538
19. Petruccelli, J.D., et al., loc. cit., p. 602
20. Devore, J.L., *Probability and Statistics for Engineering and the Sciences*, 3rd Ed., pp. 392 – 393, Brooks/Cole, Pacific Grove, CA, (1991)
21. Walpole, R.E., et al., loc. cit., Section 3.17, pp. 481-483
22. Dunnett, C.W., "A multiple comparisons procedure for comparing several treatments with control," *J. Am. Stat. Assoc.*, **50**, 1096 (1955)
23. Dunnett, C.W., "New tables for multiple comparisons with a control," *Biometrics*, **3**, 482 (1964)
24. Walpole, R.E., et al., loc. cit., Table A.13, pp. 690-691
25. Hogg, R.V., and J. Ledolter, *Engineering Statistics*, Example 5.1-2, p. 195, Macmillan, NY; Collier, London (1987)
26. Hogg, R.V., and J. Ledolter, *Applied Statistics for Engineers and Physical Scientists*, 2nd Ed., Example 7.1.2, p. 265, Macmillan, NY, (1992)
27. Walpole, R. E., et al., loc. cit., p.471
28. Beyer, W.H. (ed.), *CRC Handbook of Tables for Probability and Statistics*, 2nd Ed., Section VI.5, pp. 325-327, CRC Press, Boca Raton, FL, (1981)
29. Kokoska, S., et al., loc. cit., Table 16, p.74
30. Guenther, W. C., *Analysis of Variance*, Section 1-13, pp. 21-22, Prentice – Hall, Englewood Cliffs, NJ, (1964)
31. Walpole, R.E., et al., loc. cit., Table A.11, pp.686-687
32. Guenther, W.C., loc. cit., Table 5, pp. 184-185
33. Guenther, W.C., loc. cit., Section 1-12, pp. 20-21
34. Walpole, R. E., *Introduction to Statistics*, Section 12.3, pp. 299-300, Macmillan, NY; Collier – Macmillan, London (1968)
35. Walpole, R.E., et al., loc. cit., Section 13.4, pp. 469-471
36. Walpole, R.E., et al., loc. cit., Table A.10, pp. 684-685
37. Kokoska, S., et al., loc. cit., Table 15, pp. 72-73

APPENDIX 1. TESTING THE HYPOTHESIS OF EQUAL-POPULATION VARIANCES

The null hypothesis

$$H_0:\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2 \qquad (A.1)$$

is a statement of homoscedasticity, or homogeneity of variances, against the alternative hypothesis H_a: at least two variances are unequal.

(i) Equal-size observations

The Cochran test[27-31] compares the numerical value of the G – statistic

$$G=\frac{\max_{(i)} S_i^2}{\sum_{i=1}^{k} S_i^2} \qquad (A.2)$$

composed of the sample variances, to critical values of $g(\alpha;k,n)$ tabulated for $\alpha = 0.05$ and 0.01; H_0 in Eq. (A.1) is rejected if $G > g(\alpha;k,n)$. In our case g = 11.857/(6.857+11.357+5.0714) = 0.4877 being less than g(0.05;3,8) = 0.6530,[27-32] we fail to reject the assumption of variance homogeneity among the three sets of conversion.

(ii) Unequal-size observations

Two versions of the Bartlett test for Eq. (A.1) are given in the literature. In the earlier version[33,34] of the test a pooled variance estimate

$$S_p^2=\frac{\sum_{i=1}^{k}\left(n_i-1\right)^2 S_i^2}{N-k} \qquad (A.3)$$

is used to compute the Bartlett – statistic

$$B_1=\frac{M}{C}\ln(10) \qquad (A.4)$$

where

$$M=\left(N-k\right)\log_{10}\left(S_p^2\right)-\sum_{i=1}^{k}\left(n_i-1\right)\log_{10}\left(S_i^2\right) \qquad (A.5)$$

and

$$C=1+\frac{1}{3\left(k-1\right)}\left(\sum_{i=1}^{k}\frac{1}{n_i-k}-\frac{1}{N-k}\right) \qquad (A.6)$$

The B_1 – statistic is approximately chi – square distributed with (k – 1) degrees of freedom.

In the more recent version,[35] the pooled variance computed via Eq. (A.3) is used to obtain the B_2 – statistic defined as

$$B_2=\frac{\left[\prod_{i=1}^{k}\left(S_i^2\right)^{n_i-1}\right]^{1/\left(N-k\right)}}{S_p^2} \qquad (A.7)$$

belonging to the Bartlett – distribution. It has the curious, if not confusing, distinction that H_0 is rejected if B_2 has a numerical value less than the composite critical value

$$b_k\left(\alpha;n_1,n_2,\ldots,n_k\right)=\frac{\sum_{i=1}^{k} n_i b_k\left(\alpha;n_i\right)}{N} \qquad (A.8)$$

For our catalysts, $S_1^2 = 6.857$; $S_2^2 = 8$; $S_3^2 = 2.8$, hence $S_p^2 = 6$; M = 0.62441; C = 1.0777, and $b_1 = 1.334$. At $\alpha = 0.05$, the critical chi – square value at DF = 2 is 5.991 (found in any statistics textbook or tables), and we fail to reject H_0 in Eq. (A.1). In the second version, B_2 has the numerical value of 0.919 from Eq. (A.7). The critical value at $\alpha = 0.05$ is computed[36,37] as $b_3(0.05;8,6,6) = [8(0.7387)+2(6)(0.6483)]/20 = 0.6845$ via Eq. (A.8). Since (*careful!*) 0.919 is *larger* than 0.6845, we fail to reject the null hypothesis of equal population variances(at $\alpha = 0.01$, the critical value, [8(0.6282) + 2(6)(0.5149)]/20 = 0.5602 is expectedly lower).

APPENDIX 2. ELEMENTARY CONCEPTS OF CONTRAST THEORY

Suppose we have tested a larger number—say, five—catalysts in our reactor and have at our disposal five observation means $\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_5$. We may want to test more involved null hypotheses, *e.g.*, H_0: $(\mu_1-\mu_2)-(\mu_3-\mu_4)=0$; H_0: $(\mu_1-\mu_3)-(\mu_2-\mu_5)=0$, etc. for some reason. These combinations are called contrasts. In general, a contrast is $c_1\mu_1+c_2\mu_2+\ldots+c_m\mu_m$ where the c_i coefficients can be positive, negative, or zero. The confidence intervals are accordingly more complicated than for the ones shown in the text. For example, if we employ the Scheffé method, the expression

$$\sum_{i=1}^{m} c_i x_{i\cdot}-\left[\sum_{i=1}^{m}\frac{c_i^2}{n_i}\right]^{1/2}\Psi\left(\alpha;n,m\right)\prec\sum_{i=1}^{m} c_i\mu_i\prec\sum_{i=1}^{m} c_i x_{i\cdot}+\left[\sum_{i=1}^{m}\frac{c_i^2}{n_i}\right]^{1/2}\Psi\left(\alpha;n,m\right)$$

$$\Psi\left(\alpha;n,m\right)\equiv\sqrt{\left(n-1\right)\left(MSE\right)F\left(\alpha;n-1,\left(n-m\right)\right)} \qquad (A.9)$$

provides the α-level confidence interval in the case of unequal-size observation sets. If m =2, Eq. (5) is regained.

For the sake of illustration, we assume that four catalysts with observation set sizes $n_1 = 10$; $n_2 = 8$; $n_3 = 7$; $n_4 = 9$ yield mean conversions 72; 75;80;82%, respectively, and MSE =195.3 . We propose to set the null hypothesis that $(\mu_4-\mu_2)-(\mu_3-\mu_1)=0$. Here, $c_1 = 1$; $c_2 = -1$; $c_3 = -1$ and $c_4 = 1$. Also, n – m = 34 – 4 =30. Accordingly, $c_1x_{1\cdot}+c_2x_{2\cdot}+c_3x_{3\cdot}+c_4x_{4\cdot}=-1$, and $(\sum c_i^2/n_i)^{1/2} = 0.6921$. Since f[0.05;3,30]) = 2.92, Eq. (A.9) yields LB = –29.627; UB = 27.627, and we fail to reject the null hypothesis (P-value ≈ 0.004). ❐