A Ranking Test in the Biological Sciences

KENNETH MULLEN

Department of Biometry, Medical College of Virginia, Richmond

Laboratory analyses of biological materials are ranked in order of magnitude and summed across materials to give a list of laboratory scores. Under the assumed hypothesis that there is in fact no difference between laboratories, Monte-Carlo techniques are used to establish two-tailed 5% rejection limits for various combinations of laboratories and materials. The hypothesis that there is no difference between laboratories is rejected if any laboratory's score lies outside the 5% limits.

Suppose that one needs to run a group of tests on a particular set of materials (chemical or biological), using a number of different laboratories, and wishes to insure before starting that the laboratories are reliable, i.e., that (a) they run the test according to required specifications or directions and (b) if they run the same test twice, they will get, within some tolerated instrument variation, the same results.

I shall develop a statistical test here based on the ranked laboratory results which does not assume that the data have any particular distribution. The basis for this work was done by Dr. W. J. Youden of the National Bureau of Standards (1963), but his work is done primarily with a view to industrial applications.

I have endeavored here to simplify the statistical procedures and to stress biological applications by way of examples.

Experimental

Suppose that we have a number of different materials, A, B, C, . . . which we want to be analyzed by a number of different laboratories, 1, 2, 3, . . . Consider material A which will be analyzed quantitatively by all the laboratories; rank the results as follows: give the rank *1* to that laboratory with the highest numerical result. Give the rank *2* to the laboratory with the next highest result, and so on until the laboratory with the lowest result gets the highest rank. Repeat this same procedure for all the materials and record the results in a format similar to table 1.

Example 1: There are seven protein materials to be analyzed by 15 different laboratories. Each material is analyzed quantitatively, and the laboratory with the highest result is given the rank *1*, the next *2*, and so on. The results are shown in figure 2.

If a tie exists, e.g. two laboratories tie for third place, assign the rank of 3.5 to each. If three are tied for fourth place, assign the middle rank (here equal to 5) to all three. The ranks are then added across (i.e., a sum of ranks, or a score, is found for each laboratory).

It is clear that the minimum possible score is 7 (highest every time) and the maximum score is 105 (lowest every time). The average score (just halfway between the maximum and minimum) is 56.

The obvious question that one should ask of the data is, "On the basis of the data shown, how can I judge which laboratories are consistently too high or too low in their analysis?" It is clear that laboratory 4 has a higher score than the others, and laboratory 13 has the lowest score, but are these differences attributable to physical reasons, i.e., faulty analysis, or are they merely due to natural (or random) variation? After all, one laboratory must be first and one must be last.

Statistical Analysis

The development of the theory is as follows. If no real reason exists for one laboratory to be higher or lower than the others, then the ranks of the laboratories for a particular material are random variables, and each laboratory is equally likely to get a particular rank. Furthermore, the scores of the laboratories will cluster around some central point (the mean). To find what kind of result might be obtained if in fact the laboratories are not really different, I will simulate the scoring procedure with a method known to be random, viz. take 15 cards numbered 1 through 15, shuffle them into a random order. Taking the top card, write its number against laboratory 1, write the next card's number against laboratory 2, and so on until all 15 cards have been viewed. Repeat this until seven sets of ranks have

TABLE 1

		Materials				
		A	B	C	D	
Laboratories	1					
	2					
	3					
	4					
	.					
	.					
	.					

MCV QUARTERLY 2(3): 212-214, 1966

been entered for each laboratory, then sum the ranks and obtain a score for each laboratory. If this process is repeated a large number of times, one will have an idea of how the scores are distributed.

With the aid of a computer, this process was repeated 1,000 times, giving 1,000 sets of 15 laboratory scores, or 15,000 scores. Examination of these scores showed that there were 26 scores of 23 or less and 24 scores of 89 or more, making a total of 50 lying outside the limits 23 to 89. These 50 make about one-third of 1% of the 15,000. There are 15 scores (laboratories) in any particular test, so that the chance of a given test having one of these extreme scores is 15 times ⅓%, or 5%.

Remember that the basic assumption was that no laboratory was different from any other, and that differences in scores were due to random variation. Thus, we can say that if there is no underlying reason for laboratories to differ, then 5% of the time we will have scores outside the range 23 to 89. Thus if one uses the limits 23 and 89 as a criterion for judging the laboratories, and there is no difference between laboratories, one will record a difference (i.e., make a mistake) one time in twenty.

Using these criteria, from table 2 one can conclude that laboratory 4 has results which are not due to random variation, i.e., there is some physical reason for laboratory 4's high score.

The numbers 23 and 89 are called the 5% limits in the case of 15 laboratories and seven materials. It would be convenient to have similar limits for various combinations of materials and laboratories (table 3).

If, more generally, we have L laboratories and M materials, then the sum of the ranks for any laboratory is $(1 + 2 + 3 + \ldots + L)$, or $L(L + 1)/2$, and the mean sum of the rank for each laboratory is the sum of the ranks divided by L, i.e., $(L + 1)/2$; thus, the mean score for the sum of M rankings is $M(L + 1)/2$.

When no real differences exist between the laboratories, the expected sum of squares about the mean, of the scores for M rankings is

$$S^2 = M \sum_{i=1}^{L} \left\{ i - \frac{[L + 1]}{2} \right\}^2 = \frac{ML(L^2 - 1)}{12} \qquad (1)$$

Denote by S'^2 the actual sum of squares about the mean of the scores for each laboratory. Friedman (1937) has shown that

$$\chi^2_{L-1} = \frac{S'^2}{S^2} \qquad (L - 1) \qquad (2)$$

is distributed approximately as a χ^2 variate with $L - 1$ degrees of freedom.

If there are no differences between laboratories, then S'^2/S^2 should be close to unity, and if differences do exist, S'^2/S^2 would be greater than unity. Thus, we compare χ^2_{L-1} to the χ^2 variate with $L - 1$ degrees of freedom and reject the hypothesis that no real differences exist for large values of χ^2_{L-1}.

From the figures in table 2, $S^2 = 1960$, $S'^2 = 3030$, $\chi^2_{L-1} = \chi^2_{14} = 21.56$. Since the 90% limit for a χ^2 variate with 14 degrees of freedom is 21.06, we reject the hypothesis of all laboratories being the same at the 10% level.

In summary, formula (2) provides a quick method for evaluating the data to see whether differences between laboratories exist. To find which is the offending laboratory, use of table 3 is required.

Example 2: Suppose that a number of volunteers is required for a breathing experiment, and that for some reason or other it is necessary to accept only those whose duration of apnea is average; neither too long nor too short.

If there are 14 volunteers, one might check the duration of apnea three times each. Thus measure the duration of apnea for each of the 14, rank them giving the one with the longest duration the rank *1*, the next *2*, and so on until the volunteer with the shortest duration is given the rank *14*. Repeat this three times and sum the ranks for each volunteer. Table 3, with the vertical column at 14, and the horizontal at 3, gives the 5% limits, 4 and

TABLE 2

Laboratory	Materials Analyzed							
	A	B	C	D	E	F	G	Sum
1	8.0	4.0	11.5	12.0	1.5	1.0	13.5	51.5
2	15.0	15.0	1.0	4.0	15.0	15.0	1.0	66.0
3	7.0	9.0	15.0	6.0	5.0	10.0	2.0	54.0
4	14.0	13.0	14.0	15.0	13.0	14.0	9.0	92.0
5	11.5	8.0	8.5	3.0	5.0	8.0	3.0	47.0
6	6.0	2.5	6.5	13.5	9.5	11.0	10.0	59.0
7	3.0	5.5	13.0	1.0	7.0	13.0	12.0	54.5
8	11.5	10.0	11.5	13.5	14.0	12.0	5.0	77.5
9	4.5	7.0	4.5	8.5	5.0	5.0	13.5	48.0
10	2.0	2.5	8.5	2.0	3.0	6.5	11.0	35.5
11	4.5	11.5	3.0	10.0	1.5	2.0	15.0	47.5
12	1.0	1.0	2.0	6.0	9.5	3.0	7.0	29.5
13	9.0	5.5	4.5	11.0	8.0	4.0	6.0	48.0
14	11.5	14.0	10.0	8.5	11.0	6.5	4.0	65.5
15	11.5	11.5	6.5	6.0	12.0	9.0	8.0	64.5
	Average = 56							

41. Thus if any volunteer's score falls outside of these limits, reject him as being non-average.

Example 3: If, instead of testing for a difference between laboratories, one wishes to test for a difference between for example 10 chemical analyzing machines using six different substances, then the routine is as before. Rank each substance by machine and add the totals for each machine. From table 3, the 5% limits are 14 and 52.

Discussion

It would appear more profitable at first glance, to leave the laboratory analysis results in their raw state, rather than ranking them, and to perform a straightforward ANOVA. However, such an analysis would have to assume the underlying normality of the data and would at the same time not have the advantages of simplicity inherent in this design. This ranking test appears to be a useful tool for the statistically unsophisticated to determine departures from "averageness."

References

FRIEDMAN, H. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32: 675–701, 1937.

YOUDEN, W. J. Ranking laboratories by round-robin tests. *Mater. Res. Stand.* 3 (1): 9–13, 1963.

TABLE 3
Approximate 5% Probability Limits for Ranking Scores

Number of Laboratories Participating	Number of Materials												
	3	4	5	6	7	8	9	10	11	12	13	14	15
3		4	5	7	8	10	12	13	15	17	19	20	22
		12	15	17	20	22	24	27	29	31	33	36	38
4		4	6	8	10	12	14	16	18	20	22	24	26
		16	19	22	25	28	31	34	37	40	43	46	49
5		5	7	9	11	13	16	18	21	23	26	28	31
		19	23	27	31	35	38	42	45	49	52	56	59
6	3	5	7	10	12	15	18	21	23	26	29	32	35
	18	23	28	32	37	41	45	49	54	58	62	66	70
7	3	5	8	11	14	17	20	23	26	29	32	36	39
	21	27	32	37	42	47	52	57	62	67	72	76	81
8	3	6	9	12	15	18	22	25	29	32	36	39	43
	24	30	36	42	48	54	59	65	70	76	81	87	92
9	3	6	9	13	16	20	24	27	31	35	39	43	47
	27	34	41	47	54	60	66	73	79	85	91	97	103
10	4	7	10	14	17	21	26	30	34	38	43	47	51
	29	37	45	52	60	67	73	80	87	94	100	107	114
11	4	7	11	15	19	23	27	32	36	41	46	51	55
	32	41	49	57	65	73	81	88	96	103	110	117	125
12	4	7	11	15	20	24	29	34	39	44	49	54	59
	35	45	54	63	71	80	88	96	104	112	120	128	136
13	4	8	12	16	21	26	31	36	42	47	52	58	63
	38	48	58	68	77	86	95	104	112	121	130	138	147
14	4	8	12	17	22	27	33	38	44	50	56	61	67
	41	52	63	73	83	93	102	112	121	130	139	149	158
15	4	8	13	18	23	29	35	41	47	53	59	65	71
	44	56	67	78	89	99	109	119	129	139	149	159	169