A k-SAMPLE SLIPPAGE TEST FOR AN EXTREME POPULATION

By Frederick Mosteller

Harvard University

1. Summary. A test is proposed for deciding whether one of k populations has slipped to the right of the rest, under the null hypothesis that all populations are continuous and identical. The procedure is to pick the sample with the largest observation, and to count the number of observations r in it which exceed all observations of all other samples. If all samples are of the same size n, n large, the probability of getting r or more such observations, when the null hypothesis is true, is about k^{1-r}.

Some remarks are made about kinds of errors in testing hypotheses.

2. Introduction. The purpose of this paper is to describe a significance test connected with a statistical question called by the present author "the problem of the greatest one." Suppose there are several continuous populations $f(x - a_1)$, $f(x - a_2), \cdots, f(x - a_k)$, which are identical except for rigid translations or slippages. Suppose further that the form of the populations and the values of the a_i are unknown. Then on the basis of samples from the k populations we may wish to test the hypothesis that some population has slipped further to the right, say, than any other. In other words, we may ask whether there exists an $a_i > \max(a_1, a_2, \cdots, a_{i-1}, a_{i+1}, \cdots, a_k)$. From the point of view of testing hypotheses, the existence of such an a_i is taken to be the alternative hypothesis. A significance test will depend also on the null hypothesis. We shall take as the null hypothesis the assumption that all the a's are equal: $a_1 = a_2 = \cdots = a_k$.

Using these assumptions it is possible to obtain parameter-free significance tests that some population has a larger location parameter (mean, median, quantile, say) than any of the other populations.

The problem of the greatest one is of considerable practical importance. Among several processes, techniques, or therapies of approximately equal cost, we often wish to pick out the best one as measured by some characteristic. Furthermore, we often wish to make a test of the significance of one of the methods against the others after noticing that on the basis of the sample values, a particular method seems to be best. The test provided in this paper allows an opportunity for inspection of the data before applying the test of significance.

The proposed test has the advantage of being rapid and easy to apply. However, the test is probably not very powerful, and in the form presented here, the test depends on having samples of the same size from each of the several populations. The equal-sample restriction is not essential to the technique, but since no very useful way of computing the significance levels for the unequal-sample case is known to the author, it does not seem worthwhile to give the formulas. They are easy to write down.

3. The test. Suppose we have k samples of size n each. It is desired to test the alternative hypothesis that one of the populations, from which the samples were drawn, has been rigidly translated to the right relative to the remaining populations. The null hypothesis is that all the populations have the same location parameter.

The test consists in arranging the observations in all the samples from greatest to least, and observing for the sample with the largest observation, the number of observations r which exceed all the observations in the $k - 1$ other samples. If $r \geq r_0$ we accept the hypothesis that the population whose sample contains the largest observation has slipped to the right of the rest and reject the null hypothesis that all the populations are identical; instead we accept the hypothesis that the sample with the largest observation came from the population with the rightmost location parameter. If $r < r_0$, we accept the null hypothesis.

The statements just made are not quite usual for accepting and rejecting hypotheses. Classically one would merely accept or reject the hypothesis that the a_i are all equal. The statements just made seem preferable for the present purpose.

Example. The following data arranged from least to greatest indicate the difference in log reaction times of an individual and a control group to three types of words on a word-association test. The differences in log reaction times have been multiplied by 100 for convenience. Longer reaction times for the individual are positive, shorter ones are negative. Does one type of word require a shorter reaction time for the individual relative to the control group than any other?

Concrete	Abstract	Emotional
−6	−16	−6
−6	−11	−5
−5	−3	−3
−5	−2	−2
−4	−2	−1
−3	−1	0
−1	−1	1
0	1	3
0	1	5
3	1	12
9	8	13
11	10	13
12	16	15
29	20	28

Here we have $k = 3$ samples of size $n = 14$ each! We note that the Abstract column has the most negative deviation, −16, and that there are two observations in that column which are less than all the observations in the other columns. Consequently $r = 2$. Under the null hypothesis the probability of ob-

taining 2 or more observations in one column less than all the observations in the others is about .33, so the null hypothesis is not rejected.

4. Derivation of test. Suppose we have k samples of size n, all drawn from the same continuous distribution function $f(x)$. Arranging observations within samples in order of magnitude the samples O_i are: $O_1 : x_{11}, x_{12}, \cdots, x_{1n}$; $O_2 : x_{21}, x_{22}, \cdots, x_{2n}$; $\cdots$; $O_k : x_{k1}, x_{k2}, \cdots, x_{kn}$.

If we consider some one sample O_i, separately, we can inquire about the probability that exactly r of its observations are greater than the greatest observation in the other $k - 1$ samples.

The total number of arrangements of the kn observations is

$$T = \frac{(kn)!}{(n!)^k}. \tag{1}$$

The number of ways of getting all n observations of O_i to be greater than all observations in the remaining samples is

$$N(n) = \frac{[(k-1)n]!}{(n!)^{k-1}0!}. \tag{2}$$

The number of ways of getting exactly $n - 1$ observations of O_i greater than all observations in the remaining samples is

$$N(n-1) = \frac{[(k-1)n+1]!}{(n!)^{k-1}1!} - \frac{[(k-1)n]!}{(n!)^{k-1}0!}. \tag{3}$$

More generally, the number of ways of getting exactly $r = n - u$ of O_i to be greater than all other observations in the remaining samples is

$$N(n-u) = \frac{[(k-1)n+u]!}{(n!)^{k-1}u!} - \frac{[(k-1)n+u-1]!}{(n!)^{k-1}(u-1)!}. \tag{4}$$

Therefore the number of ways of getting a run of $r = n - u$ *or more* observations in O_i greater than the rest is just

$$S(n-u) = \sum_{t=n-u}^{n} N(t) = \frac{[(k-1)n+u]!}{(n!)^{k-1}u!}. \tag{5}$$

However we do not choose our sample O_i at random or preassign it, as the demonstration has thus far supposed. Instead we choose that O_i which has the greatest observation in all the samples. This condition requires us to multiply $S(n - u)$ by the factor k. Consequently the probability that the sample with the largest observation has $r = n - u$ or more observations which exceed all observations in the other $k - 1$ samples is given by

$$P(r) = \frac{kS(r)}{T} - \frac{k(n!)}{(kn)!}\frac{(kn-r)!}{(n-r)!}. \tag{6}$$

As an incidental check we note in passing that

$$P(1) = \frac{k(n!)}{(kn)!}\frac{(kn-1)!}{(n-1)!} = \frac{kn}{kn} = 1.$$

We note that equation (6) may be rewritten as

$$P(r) = kC_{n-r}^{kn-r}/C_n^{kn}, \tag{7}$$

which is a useful form for some computations.

Table I gives the probability of observing r or more observations in the sample with the largest observation, among k samples of size n, which are more extreme in a preassigned direction than any of the observations in the remaining $k - 1$ samples.

5. Approximations. If we use Stirling's formula and approximations for $(1 + \alpha)^r$, for small values of α and r, we can write an approximation for equation (6) for large values of n with r and k fixed as follows

$$P(r) \sim \frac{1}{k^{r-1}}\left(1 - \frac{r(2r-1)(k-1)}{2kn}\right). \tag{8}$$

For very large n equation (8) yields

$$P(r) \sim \frac{1}{k^{r-1}}, \tag{9}$$

which is the value given in Table I for $n = \infty$ For many purposes the result given by equation (9) is quite adequate, as a glance at Table I will indicate.

6. Kinds of errors. In tests such as the one being considered here the classical two kinds of errors are not quite adequate to describe the situation.

As usual we may make the errors of

I) rejecting the null hypothesis when it is true,

II) accepting the null hypothesis when it is false.

But there is a third kind of error which is of interest because the present test of significance is tied up closely with the idea of making a correct decision about which distribution function has slipped furthest to the right. We may make the error of

III) correctly rejecting the null hypothesis for the wrong reason.

In other words it is possible for the null hypothesis to be false. It is also possible to reject the null hypothesis because some sample O_i has too many observations which are greater than all observations in the other samples. But the population from which some other sample say O_j is drawn is in fact the rightmost population. In this case we have committed an error of the third kind.

When we come to the power of the test under consideration we shall compute the probability that we reject the null hypothesis because the rightmost population yields a sample with too many large observations. Thus by the power of

TABLE I

Probability of one of k samples of size n each having r or more observations larger than those of the other k − 1 samples

$k = 2$

n \ r	2	3	4	5	6
3	.400	.100			
5	.444	.167	.048	.008	
7	.462	.192	.070	.021	.005
10	.474	.211	.087	.033	.011
15	.483	.224	.100	.042	.017
20	.487	.231	.106	.047	.020
25	.490	.235	.110	.050	.022
∞	.500	.250	.125	.062	.031

$k = 3$

n \ r	2	3	4	5	6
3	.250	.036			
5	.286	.066	.011	.001	
7	.300	.079	.018	.003	.0004
10	.310	.089	.023	.005	.0011
15	.318	.096	.027	.007	.0018
20	.322	.100	.030	.009	.0023
25	.324	.102	.031	.009	.0026
∞	.333	.111	.037	.012	.0041

$k = 4$

n \ r	2	3	4	5	6
3	.182	.018			
5	.211	.035	.004	.0003	
7	.222	.043	.007	.0009	.0001
10	.231	.049	.009	.0015	.0002
15	.237	.053	.011	.0022	.0004
20	.241	.056	.012	.0026	.0005
25	.242	.057	.013	.0028	.0006
∞	.250	.062	.016	.0039	.0010

$k = 5$

n \ r	2	3	4	5
3	.143	.011		
5	.167	.022	.0020	.0001
7	.177	.027	.0033	.0003
10	.184	.031	.0046	.0006
15	.189	.034	.0056	.0008
20	.192	.035	.0062	.0010
25	.194	.036	.0065	.0011
∞	.200	.040	.0080	.0016

$k = 6$

n \ r	2	3	4	5
3	.118	.007		
5	.138	.015	.0011	.0000
7	.146	.018	.0019	.0001
10	.152	.021	.0026	.0003
15	.157	.023	.0032	.0004
20	.160	.024	.0035	.0005
25	.161	.025	.0037	.0005
∞	.167	.028	.0046	.0008

this test we shall mean the probability of both correct rejection and correct choice of rightmost population, when it exists.

Errors of the third kind happen in conventional tests of differences of means, but they are usually not considered although their existence is probably recognized. It seems to the author that there may be several reasons for this among which are 1) a preoccupation on the part of mathematical statisticians with the formal questions of acceptance and rejection of null hypotheses without adequate consideration of the implications of the error of the third kind for the practical experimenter, 2) the rarity with which an error of the third kind arises in the usual tests of significance.

In passing we note further that it is possible in the present problem for both the null hypothesis and the alternative hypothesis to be false when $k > 2$. This may happen when there are, say, two identical rightmost populations, and the remaining populations are shifted to the left. An examination of Table I will give us an idea of what will happen in such a case. If $k = 4$, we use $r = 3$ as about the .05 level. If two of the populations are slipped very far to the left, while the rightmost two populations are identical, in effect $k = 2$. In this case the probability of rejecting the null hypothesis is around .2. Consequently we accept the null hypothesis about 80 per cent of the time, and reject it 20 per cent of the time under these conditions. But neither hypothesis was true.

If we carry the discussion to its ultimate conclusion we would need a fourth kind of error for these troublesome situations. There are still other kinds of errors which will not be considered here.

7. The power of the test. It is difficult to discuss the power of a non-parametric test, but in the present case it may be worthwhile to give an example or two. The reader will understand that although the test is called non-parametric, its power does depend on the distribution function.

In the case of k samples there are two extremes which might be considered for any particular form of distribution function. In Case 1, we suppose that when the alternative hypothesis is true, $k - 1$ of the populations are identical with distribution function $f(x)$, while the remaining distribution function is $f(x - a)$, $a > 0$. Case 1 may be regarded as a *lower bound* to the power of the test because for any fixed distance a between the location parameters of the rightmost population and the next rightmost population, Case 1 gives the least chance of detecting the falsity of the null hypothesis.

In Case 2, we suppose that the rightmost population is $f(x - a)$, $a > 0$ as before, that the next rightmost population is $f(x)$, and that the other $k - 2$ populations have slipped so far to the left that they make no contribution to problem of the power. This is an optimistic approach to the power because it gives an *upper bound* to the power. When $k = 2$, Case 1 and Case 2 are identical, and the power is exactly the power of the test for the particular distribution function under consideration.

Case 3 which we shall not consider deals with the situation where there is more

than one rightmost population, but the null hypothesis is false. It is connected with the fourth kind of error mentioned at the end of section 6.

Table II gives the upper and lower bound of the power of the test for $k = 3$, $r = 3, n = 3$, when the distribution is uniform and of length unity. The parameter a is the distance between the location parameter of the rightmost distribution and that of the next rightmost distribution.

In Table III we give some points on the upper and lower bounds of the power of the test for the normal distribution with unit standard deviation. The parameter a is the distance between the mean of the rightmost normal distribution and the next rightmost, measured in standard deviations. Again we use the case $k = 3$, $r = 3$, $n = 3$.

TABLE II

Power p of the test for the uniform distribution when $k = 3$, $r = 3$, $n = 3$. The distance between the midpoints of the two rightmost distributions is a

a	0	.1	.3	.5	.7	.9	1.00
Upper bound p_u	.05	.09	.23	.46	.73	.96	1.00
Lower bound p_l	.01	.03	.11	.29	.59	.93	1.00

TABLE III

Power p of the test for the unit normal when $k = 3$, $r = 3$, $n = 3$. The distance between the means of the two rightmost distributions, measured in standard deviations, is a

a	0	.5	1.0	1.5	2.0	2.5	3.0
Upper bound p_u	.05	.13	.26	.42	.58	.71	.87
Lower bound p_l	.01	.04	.14	.27	.43	.60	.80

The power of the test has been defined as the probability of correctly rejecting the null hypothesis and finding the sample from the rightmost population to be the extreme one. This raises a question about the meaning of the entries in Tables II and III under $a = 0$. When $a = 0$ there is no way to reject the null hypothesis correctly. The probabilities given are the probabilities that a randomly chosen sample will force a rejection of the null hypothesis. They represent the limit of the power function as a tends to zero. If we think of earmarking the sample from the rightmost population and of computing the probability repeatedly that that sample will have three observations larger than all the observations in the other sample, and then we let a tend to zero, this is the result we get. These values are not the significance levels. The significance level is .036.

8. Discussion. The reader may rightly feel that the solution here presented to the problem of the greatest one depends on a trick. That is, it depends intimately on the choice of the null hypothesis. Furthermore the reader may feel that the choice of $a_1 = a_2 = \cdots = a_k$ is neither an interesting null hypothesis nor one which is likely to arise in a practical situation. The author has no quarrel with this attitude. This means that there are many other approaches to this problem which are worth trying. The equal-location-parameter case is one which yields easily to non-parametric methods.

It will be noted that a useful technique has been indicated which allows one to examine the data before making the significance test. In general one may wish to set up a test function, decide which of several samples provides the extreme value of the function, and then test significance given that we have chosen that sample which maximizes the function among the k samples under consideration.

9. Conclusion. There is a large class of problems grouped around "the problem of the greatest one". First it would be useful to have a more powerful test than the one here proposed. Second, there is the problem of deciding on the basis of samples whether we have successfully predicted the order of the location parameters of several populations. Third, there is the general problem of what alternatives, what null hypotheses, and what test functions to use in treating samples from more than two populations. It is to be hoped that more material on these problems will appear, because answers to these questions are urgently needed in practical problems.