THE MEDIAN IN TESTS BY RANDOMIZATION

K. R. NAIR

Statistical Laboratory, Calcutta.

INTRODUCTION

Many of the statistical tests now in use are based on the assumption of normality in the parent population, from which the observed samples are supposed to be drawn. Though it is known that many of these tests remain valid even with moderate departures from normality, it is necessary to develop exact tests of significance which are free from the assumption of normality.

Recently a number of papers has been published on such tests of significance. W. R. Thompson[9] and S. R. Savur[8] by working on the "Confidence Range" of the median from any population have suggested methods for testing significance for two samples. The use of ranks has been suggested by Hotelling and Pabst[3] in correlation study and is advocated in analysis of variance by Friedman[2]. Last but not least R. A. Fisher gave an example in Section 21 of his book *"Design of Experiments"* illustrating a method based on the principle of randomization, replacing the usual t-test which is appropriate for normal samples. This idea of Fisher has been extended by Pitman[5,6] to test correlation co-efficients and by Pitman[7] and B. L. Welch [10,11] to problems arising in randomized block and Latin square designs where the principle of randomization is usually followed.

Fisher's primary object in developing his test by Randomization, in which the assumption of normality can be avoided, was to test the significance of two paired or two independent samples of equal size. In the former case he takes the modular differences $|d_1|$, $|d_2|$, ... $|d_n|$ of the n pairs of readings and as, in the absence of any treatment effect, the only values possible for any difference by randomization being plus or minus the given magnitude, he gets 2^n elements of a sub-population ; he then uses the mean-statistic of these elements for testing the significance. In the second case[1] there are two independent samples $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$. He pools the two samples and gets the $(2n)!/(n!)^2$ possible pairs of samples, and uses the difference of the mean-statistic of each pair of samples.

It is clear that the distribution of the mean in both these cases has to be done by laborious numerical work. Also, E. S. Pearson[4] has shown that this test will control the second kind of error only if the statistic chosen is efficient. But as our chief object is to avoid any assumption as to the form of the population, any discussion of efficiency of the statistic used is meaningless. Instead of using the mean-statistic it is of some practical importance to search for statistics whose distributions in the sub-population obtained by the use of the principle of randomization are easily derivable.

In this paper are obtained the distribution of the p-th ranked individual in the first type of problem and of the difference between the p-th ranked individuals in the second type of problem. Some other distributions are also considered. Incidentally, some light is thrown on Pearson's point of view, namely, that even in the test by randomization the statistic used should be efficient in order to control the second kind of error.

TEST BY RANDOMIZATION FOR TWO PAIRED SAMPLES

Consider a randomized block experiment with (say) two varieties A and B, and n replications. Let a_i and b_i be the yields of the A and B plots in the i-th block. Let us start with the hypothesis of "no difference" in the yield of the two varieties. In this case the observed yields in the two plots might have been the same even if the varieties were interchanged. The (A−B) difference in the i-th block can therefore have either of the two equally likely values (a_i-b_i) or (b_i-a_i). This is true independently for the remaining blocks. Let us denote the numerical magnitudes of the differences of the two plot yields by $|d_1|, |d_2|, \ldots \ldots |d_i|, \ldots \ldots |d_n|$. The A−B difference in the i-th block may be either d_i or $-d_i$. By randomization therefore the set of (A−B) differences of a single arrangement can have different values. We may call this a sub-population of 2^n elements.

In an actual experiment only one of these elements is obtained. The other elements are generated with the knowledge of the numerical values of the d's. If now we calculate the mean of each of the 2^n elements of n each, we get the distribution of the mean in this sub-population. If the mean of the element actually observed falls beyond the 5 per cent tail area of this distribution we reject our null hypothesis of "no-difference." If we get another set of $\{|d_i|\}$, we form another sub-population of 2^n elements. If in each sub-population the same rule of rejecting the hypothesis at 5 per cent level is followed, in the long run of our experience with all possible $\{|d_i|\}$ we will be controlling the wrong rejection of the hypothesis tested, to 5 per cent of cases even in the main population. Instead of the mean, any other statistic of location could be tested on this principle, say, median, midpoint, geometric mean, etc.

Fisher has chosen the mean, for illustrating this method on Darwin's data. The calculations involved were long and tedious for his case of 2^{15} (=32768) elements. No general expression is available which can be applied whatever be the value of n and whatever the actual magnitudes of the $\{|d_i|\}$. It will be presently seen that the median is a statistic whose distribution comes out very easily for any n and any set of values $\{|d_i|\}$. Of course the median distribution differs according as n is even or odd. What is needed is only to rank the observed $\{|d_i|\}$ according to magnitude.

We can suppose without loss of generality that $|d_i|$ is the i-th in rank, in ascending order of magnitude. Let also the n differences be distinct avoiding all 'tied-ranks'. In the 2^n elements that can be got by assigning a *plus* or a *minus* sign to the d's, the p-th ranked variate in an element will assume different values, the rank being fixed from smallest to largest. Let us see what the various values are that it can have and the frequency attached to each. Among the 2^n elements of the sub-population the one got by assigning the minus sign to all the values stand at one extreme and the one got by assigning the plus sign to all the values, at the other extreme. In the former case the p-th ranked variate is $-d_{n-p+1}$ and in the latter $+d_p$. The p-th ranked variate therefore is distributed in tho range $-d_{n-p+1}$ to $+d_p$, taking n+1 discrete values. The number of elements in which $+d_m$ ($1 \leqq m \leqq p$) will be the p-th ranked variate is ${}_{n-m}C_{n-p}\cdot 2^{m-1}$. The number of elements in which $-d_m$ ($1 \leqq m \leqq n-p+1$) will be p-th ranked variate is ${}_{n-m}C_{p-1}\cdot 2^{m-1}$. If n is odd and equal to (2k+1) the median becomes the (k+1)-th ranked variate so that its distribution follows from the foregoing. The median takes values in the range $-d_{k+1}$ and $+d_{k+1}$ with a symmetrical frequency distribution. The frequency with which the median coincides with $+d_m$ and $-d_m$ ($1 \leqq m \leqq k+1$) being ${}_{(2k+1-m)}C_k\cdot 2^{m-1}$.

In Fisher's example the 15 paired differences actually observed were 49, -67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60 , -48. Median of this element is + 24.

The 15 magnitudes arranged according to size are 6, 8, 14, 16, 23, 24, 28, 29, 41, 48, 49, 56, 60, 67, 75. The frequency distribution of the median in the sub-population of 2^{15} elements got by randomization is given in Table 1.

TABLE 1. FREQUENCY DISTRIBUTION OF MEDIAN OF PAIRED DIFFERENCES OF DARWIN'S DATA.

Value of Median	Frequency		Value of Median	Frequency	
−29	$2^7 =$	128	+6	$_{14}C_7.1 =$	3432
−28	$_8C_1.2^6 =$	512	+8	$_{13}C_6.2 =$	3432
−24	$_9C_2.2^5 =$	1152	+14	$_{12}C_5.2^2 =$	3168
−23	$_{10}C_3.2^4 =$	1920	+16	$_{11}C_4.2^3 =$	2640
−16	$_{11}C_4.2^3 =$	2640	+23	$_{10}C_3.2^4 =$	1920
−14	$_{12}C_5.2^2 =$	3168	+24	$_9C_2.2^5 =$	1152
−8	$_{13}C_6.2 =$	3432	+28	$_8C_1.2^6 =$	512
−6	$_{14}C_7.1 =$	3432	+29	$2^7 =$	128
			Total		32768

The probability of the median being $\geqq$ 24 in absolute magnitude is 3584/32768 or 11 per cent. From the 5 per cent level of significance the observed median accepts the null hypothesis tested, namely of "no-difference." Fisher's test using the mean gave him a probability of about 5·26 per cent of getting a mean greater than the one observed. This is just on the border of the region of acceptance of the null hypothesis.

Owing to the difference in definition of median when n is even (say 2k) its distribution in this case is not so easy as in the previous case. What we can do here is to get the frequency with which a given pair of varieties become the central ranked variates, k-th and (k+1)-th, of an element.

Let $+d_i$ and $+d_j$ be the central values of one of the elements, where i and j are in the range $1 \leqq i < j \leqq k+1$. The number of elements in which this happens is $_{n-j}C_{k-1} \cdot 2^{j-1}$. Also $-d_i$ and $-d_j$ occur with the same frequency. Now consider the frequency with which $+d_i$, $-d_j$ or $-d_i$, $+d_j$ can become central values. It is easily seen that i can have only value 1. The frequency with which $+d_1$ and $-d_j$ or $-d_1$ and $+d_j$ become central values is the same and equals $_{2k-j}C_{k-1}$, where $1 < j \leqq k+1$.

The median therefore covers a range of k(k+1) discrete values and has a symmetrical frequency distribution. We cannot however arrange the values of the median, from smallest to lowest, as this is not known in the general case. This fact brings to our notice that the distribution is more often multi-modal than not, the number of humps depending on the magnitudes in the data. For these reasons it will be better in even-sized samples to consider only the distribution of the k-th or (k+1)-th ranked individual.

The examination of the median in the case of n even, brings us to the interesting problem of the distribution of the *Range*, and incidentally, of the *Mid-point* in the general case of n even or odd. It is necessary in order to investigate these distributions, to get the frequency with which a given pair of variates can become the biggest and the smallest variates of an element. It is clear that the extreme variates will be both positive or both negative only when all the variates are positive or negative. In this case naturally $+d_1, +d_n$ or $-d_1, -d_n$ become the extreme values and the frequency of each is 1.

It is also seen that the extreme variates can be of opposite sign say d_i and $-d_j (i<j)$ only when $j=n$. The frequency of this is 2^{i-1}. Also $-d_i$ and $+d_n$ occur as extreme values with the same frequency. The frequency distribution of the *Midpoint* is therefore as in Table 2.

TABLE 2. DISTRIBUTION OF MID-POINT OF PAIRED DIFFERENCES.

Least	Greatest	Midpoint	Frequency
$-d_{n-1}$	$+d_n$	$\frac{1}{2}(d_n-d_{n-1})$	2^{n-2}
$-d_{n-2}$	$+d_n$	$\frac{1}{2}(d_n-d_{n-2})$	2^{n-3}
⋮	⋮	⋮	⋮
$-d_1$	$+d_n$	$\frac{1}{2}(d_n-d_1)$	$2^0=1$
$+d_1$	$+d_n$	$\frac{1}{2}(d_n+d_1)$	1
$-d_n$	$-d_1$	$-\frac{1}{2}(d_n+d_1)$	1
$-d_n$	$+d_1$	$-\frac{1}{2}(d_n-d_1)$	$2^0=1$
⋮	⋮	⋮	⋮
$-d_n$	$+d_{n-1}$	$-\frac{1}{2}(d_n-d_{n-1})$	2^{n-3}
$-d_n$	$+d_{n-1}$	$-\frac{1}{2}(d_n-d_{n-1})$	2^{n-2}
		Total	2^n

This is a symmetrical but U-shaped distribution. Since $-d_1, +d_n$ or $-d_n, +p_1$ as extreme variates give the same value for the range namely d_n+d_i the frequency of this value of the range is 2^i. The distribution of the *Range* is therefore J-shaped and is given in Table 3.

TABLE 3. DISTRIBUTION OF RANGE OF PAIRED DIFFERENCES.

Range	Frequency
d_n+d_{n-1}	2^{n-1}
d_n+d_{n-2}	2^{n-2}
⋮	⋮
d_n+d_1	2
d_n-d_1	2
Total	2^n

Of the various statistics whose distributions we have studied, it is apparent that the median individual in the case of an odd sample and one of the two central individuals in the case of an even sample are convenient statistics to use in tests of significance of a given set of differences based on the principle of randomization.

Test by Randomization of two equal but Independent Samples

Fisher[1] has recently given an illustration of the application of the principle of randomization to two independent samples of n observations each. Let the original samples belong to two groups A and B. Our problem is to see whether the samples give evidence of real difference between A and B with regard to centre of location (mean, median, etc.) If the $2n$ observations are lumped together they can be split into two groups of n each and assigned to A and B in $(2n)!/(n!)^2$ ways. For each of these ways we can calculate a difference between the means or medians or any other measure of centre of location, of the two groups. We thus get a sub-population of $(2n)!/(n!)^2$ differences of the statistic studied. If the difference observed in the original two samples is such that the probability of such a value or a greater value occurring in this sub-population is (say) less than 0·05 we agree to declare that the original samples reject the hypothesis of "no-difference" tested.

If we choose the mean for this test it is easy to see that the process is very lengthy in getting its distribution in the sub-population of $(2n)!/(n!)^2$ differences. It has also to be done independently on every occasion. It will be seen presently that when n is odd the distribution of the median differences comes out very easily and is general.

Let n be even or odd. Let the pooled data of the two samples arranged according to ascending order of (algebraic) magnitude be $x_1, x_2, \ldots\ldots x_{2n}$. Let us confine our attention to the p-th ranked variates of the two samples, each of size n, that can be made out of these $2n$ observations in $(2n)!/(n!)^2$ ways. Let us determine the number of sample pairs in which x_i and x_j $(i<j)$ will be the p-th ranked variates of the two samples. This can be easily seen to be twice

$$ {}_{i-1}C_{p-1} \cdot {}_{j-i-1}C_{j-2p} \cdot {}_{2n-j}C_{n-p}, $$

i and j being governed by the following inequalities

$$ j > i \geqq p $$
$$ 2p \leqq j \leqq n+p. $$

We have thus obtained the frequency distribution of the quantity $|x_i - x_j|$ occurring as the difference between the p-th ranked variates of the two samples.

From this by putting $n=2m+1$ and $p=m+1$ we get the distribution of the differences between the medians of two samples having equal but odd number of observations. The frequency of x_i and x_j becoming the medians, thus giving a median difference $|x_i - x_j|$ is twice

$$ {}_{i-1}C_{m} \cdot {}_{j-i-1}C_{j-2m-2} \cdot {}_{4m+2-j}C_{m}. $$

If $n=2m$, the distribution of the median (defined as mean of the two central values) is rather inconvenient to use. In that case it is perhaps good enough to study the distribution of either of the two central values got by putting $p=m$ or $m+1$ in the above distribution.

This expression completely specifies the frequency distribution of the median differences. Only, we cannot arrange the frequencies according to magnitude of the differences, as the latter is not known in gneral. For a known set of observations this can be done easily.

E. S. PEARSON'S PROBLEM

I shall illustrate the distribution of the median differences on one of the examples given by E. S. Pearson.[4] He took 5 sets of two samples of 7 individuals in each from two rectangular populations with *different* centres of location but having the same range. They are reproduced in Table 4.

TABLE 4. E. S. PEARSON'S EXPERIMENTAL SAMPLING DATA.

Experiment	I		II		III		IV		V	
Sample.	1	2	1	2	1	2	1	2	1	2
	45	120	29	50	14	60	47	60	67	47
	21	122	41	125	70	104	4	90	18	71
	69	107	27	112	32	81	49	84	41	43
	82	127	5	86	79	41	49	100	41	115
	79	124	27	40	87	69	23	93	65	66
	93	41	58	98	25	40	52	32	8	124
	34	37	92	50	2	48	67	98	52	56
Midpoint	57·00	82·00	48·50	82·50	44·50	72·00	35·50	66·00	37·50	83·50
Mean	60·43	96·86	39·86	80·14	44·14	63·29	41·57	79·57	41·71	74·57
Median	69·00	120·00	29·00	86·00	32·00	60·00	49·00	90·00	41·00	66·00

He has virtually worked out the distribution of the differences of the mean and of the midpoint and shows that the test based on the midpoint rejects the hypothesis of "no-difference" (which is definitely known to be wrong) more often than the test based on the mean. This is because for rectangular population the midpoint is an efficient statistic. In samples of $n(=2m+1)$ from a rectangular law $0 \leqq x \leqq b$ the variance of the midpoint, mean and median are respectively

$$\frac{b^2}{4(m+1)(2m+3)}, \quad \frac{b^2}{12(2m+1)}, \quad \frac{b^2}{4(2m+3)}$$

Since the median has the biggest variance its efficiency is less than that of the mean. Accordingly if in Pearson's examples we apply the median test we should expect according to Pearson's contention that it will accept the wrong hypothesis we are testing, more often than either the midpoint or the mean. I have performed the test on all the 5 sets of two samples, of 7 each, he has used and have got results which confirm his arguments.

The distribution of median differences for the first set used by E. S. Pearson is given in Table 5. The 14 observations according to size are 21, 34, 37, 41, 45, 69, 79, 82, 93, 107, 120, 122, 124, 127.

The number of pairs of samples under randomization having a median difference equal to or greater than the difference in the observed pair is 552. The corresponding frequencies for difference of midpoint and mean given by E. S. Pearson are 90 and 252 respectively. Thus the median accepts the wrong hypothesis in 552 cases while the mean and midpoint accept in 252 and 90 cases respectively. The median accepts the wrong hypothesis more often than the other two statistics.

TABLE 5. FREQUENCY DISTRIBUTION OF MEDIAN DIFFERENCE OF PEARSON'S EXPERIMENT I.

Value of Median				Median Difference	Frequency
Sample A (B)		Sample B (A)			
i	x_i	j	x_j	$x_j - x_i$	
4	41	11	120	79	2 . C = 40
4	41	10	107	66	2 . C . C = 80
4	41	9	93	52	2 . C . C = 80
4	41	8	82	41	2 . C = 40
5	45	11	120	75	2 . C . C = 80
5	45	10	107	62	2 . C . C . C = 192
5	45	9	93	48	2 . C . C . C = 240
5	45	8	82	37	2 . C . C = 160
6	69	11	120	51	2 . C . C = 80
6	69	10	107	38	2 . C . C . C = 240
6	69	9	93	24	2 . C . C . C = 400
6	69	8	82	13	2 . C . C = 400
7	79	11	120	41	2. C = 40
7	79	10	107	28	2 . C . C = 160
7	79	9	93	14	2 . C . C = 400
7	79	8	82	3	2 . C . C = 800
				Total	3432

Similar results have been worked out for all the 5 experiments and are summarized in Table (6). The results of the mean and midpoint are taken from E. S. Pearson's paper.

TABLE 6. NUMBER OF PAIRS OF SAMPLES, UNDER RANDOMIZATION, HAVING VALUES FOR DIFFERENCES BETWEEN MID-POINT, MEAN, AND MEDIAN AS GREAT AS, OR GREATER THAN THE OBSERVED PAIRS (E. S. PEARSON'S DATA).

Experiment	Midpoint	Mean	Median
I	90	252	552
II	108	114	120
III	282	>506	1512
IV	62	40	200
V	106	168	120

For all the experiments we see that the test using midpoint as the centre of location rejects the (wrong) hypothesis we are testing more often than those using mean or median.

References

1. Fisher, R. A.: *J. R. Anthrop. Inst.* Vol. LXVI (1936), p. 57.

2. Friedman, M.: The Use of Ranks to avoid the assumption of Normality. *J. Amer. St. Assoc.*, Vol. 32 (1937), pp. 675-701.

3. Hotelling, H. and Pabst, M. R.: Rank Correlation and Tests of Significance Involving no Assumption of Normality. *Ann. Math. Stat.*, Vol. VII (1936), pp. 29-43.

4. Pearson, E. S.: Some Aspects of the Problem of Randomization. *Biometrika*, Vol. XXIX (1937), pp. 53-64.

5. Pitman, E. J. G.: "Significance Tests which may be applied to Samples from any Populations". *Suppl. J. R. S. S.*, Vol. IV (1). (1937), pp. 119-130.

6. Pitman, E. J. G.: Significance Tests which may be applied to Samples from any Population. III. The Correlation Coefficient Test. *Suppl. J. R. S. S.*, Vol. IV (2). (1937), pp. 225-232.

7. Pitman, E. J. G.: Significance Tests which may be applied to Samples from any Population. III. The Analysis of Variance Test. *Biometrika*, Vol. XXIX (1937), pp. 232-235.

8. Savur, S. R.: The Use of the Median in Tests of Significance. *Proc. Ind. Acad. Sc.*, (Section A) Vol. V (1937), pp. 564-576.

9. Thompson, W. R.: On Confidence Ranges for the Median and other Expectation Distributions for Population of Unknown Distribution Form. *Ann. Math. Stat.*, Vol. VII (1936), pp. 122-128.

10. Welch, B. L.: On the z-test in Randomized Blocks and Latin Squares. *Biometrika*, Vol. XXIX (1937), pp. 21-52.

11. Welch, B. L.: On Tests for Homogeneity. *Biometrika*, Vol. XXX (1938), pp. 149-158.