

**Experimental Evaluation of the Efficiencies  
of Certain Non-Parametric Statistics**

**By**

**Frederick W. Cleaver**

**A Thesis submitted to the Graduate Committee  
for the Degree of  
Master of Science  
in  
Statistics**

**Approved:**

\_\_\_\_\_  
**Head of Department**

\_\_\_\_\_  
**Chairman, Graduate Committee**

\_\_\_\_\_  
**Dean of College**

**Virginia Polytechnic Institute**

**June, 1950**

## Table of Contents

	Page
I. Introduction-----	1
II. Purpose-----	7
III. Method-----	9
IV. Results-----	16
V. Discussion of Results-----	25
VI. Bibliography-----	27

## I. Introduction

Many of the earlier developments in the field of statistics were formulated under the assumption that the random variables under investigation were distributed normally with c. d. f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Many tests and analysis of variance techniques were derived for evaluation of data from normal populations, and, in fact, transformations have been found which when applied to some classes of non-normal data permit the use of these tests and techniques. The development and adaptation of these techniques and tests makes possible the drawing of inferences from the observed data if the observer can be sure, or reasonably sure, that the random observations come from populations which are distributed according to one of the known distribution functions. The observer will be more likely to accept inferences drawn from data that very closely approximates one of the known distributions, because he will feel more certain that his computed statistics closely approximate the parameters which characterize the distribution. When the observer can be reasonably sure that the observed data reveals the parameter in question, he may make inferences with regard to whether or not his observed estimates of the population parameters are significantly different from the population parameters which would exist under the particular hypothesis in question. Such inferences may be called "Parametric Inferences". The making of such inferences is the most common activity of statisticians.

The property of credibility, which is possessed to a greater or lesser extent by all statistical inferences, is defined as the relation of protasis to apodosis, or the relation of the set of data, hypothesis, evidence, etc. to the set of conclusion.<sup>(1)</sup> One of the most important factors entering into the credibility which will be attached to a particular inference is the amount of reliance that may be placed on the assumption that the observed statistics are calculated from data gathered from a population which is distributed according to that distribution function which is assumed to characterize the distribution of the data and which provides the parameters which the calculated and observed statistics are believed to represent. When it is possible to place high reliance on such an assumption because of prior knowledge of the general characteristics of the population under investigation, the credibility that will be placed on inferences drawn from relationships among the observed statistics and their corresponding hypothicated parameters will be high. Thus, it may be seen that in parametric inference, high credibility will be attached to inferences when the assumptions as to the distribution of the underlying population are justified. The degree to which these assumptions are justified determines the intensity of the credibility that the inference will enjoy.

What, then, can the experimenter do if he cannot justify the assumption that the underlying population under investigation is such that it can be characterized by any of the previously determined distribution functions? Must he choose a distribution

function more or less at random and thereby accept the consequent diminution in the credibility that may be attached to his results? Or is it possible to evolve new statistics, statistics which do not rely for their credibility on any particular underlying distribution function, statistics which are not estimates of parameters, in short, non-Parametric Statistics?

Non-Parametric statistics have been evolved for the use of the experimenter who finds himself in such a position. Prominent among the techniques in the field of non-parametric inference is the use of rank order statistics. Three particular rank order tests are to be investigated in this thesis. They are the Wald-Wolfowitz "U"(2), the Wilcoxon "T"(3), and the Mann-Whitney "U"(4).

The Wald-Wolfowitz test utilizes a statistic "U" which is defined by considering a sub-sequence  $v_{s+1}, v_{s+2}, \dots, v_{s+r}$  of  $V$  where  $V = v_1, v_2, \dots, v_{m+n}$ , a sequence defined, as follows:  $v_i = 0$  if  $Z_i$  is a member of the set  $X_1, X_2, \dots, X_m$  and  $v_i = 1$  if  $Z_i$  is a member of the set  $Y_1, Y_2, \dots, Y_n$  where the set of  $Z$ 's is the sequence composed of the numbers in the two sets,  $X_1, X_2, \dots, X_m$ , and  $Y_1, Y_2, \dots, Y_n$  arranged in ascending order of magnitude. In this sub-sequence of  $V$  (where  $r$  may also be 1) a "run" shall exist if  $v_{s+1} = v_{s+2} = \dots = v_{s+r}$  and if  $v_s \neq v_{s+1}$  where  $S = 0$  and if  $v_{s+r} \neq v_{s+r+1}$  when  $s+r < m+n$ . "U" is then defined as the number of runs in "V" and is used to test the hypothesis that  $f(x) \cong g(x)$ . A difference between  $f(x)$  and  $g(x)$  decreases the number of runs and hence decreases "U". The distribution of "U" under the null-hypothesis has been evaluated by Wald and Wolfowitz

in their paper. The asymptotic distribution of "U" has been shown by them to converge uniformly in "t" to

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{u^2}{2}} du$$

when  $n \rightarrow \infty$ , where  $\frac{m}{n} \rightarrow c$ , a positive constant such that

$$E(U) \sim \frac{2m}{1+c} \quad \text{and} \quad \sigma^2(U) \sim \frac{4cm}{(1+c)^2}$$

and where "t" is any real number. The consistency of this test (i.e., the probability of rejecting the null-hypothesis when it is false) has been shown to approach 1 as a limit when n approaches infinity. This does not shed any light on the consistency of the test for small n.

The Wilcoxon test (which is applicable only to cases in which the same number of observations are made from both populations) assigns ranks to the experimental results in ascending order in the case of unpaired comparisons and to the differences in experimental results in the case of paired comparisons. The test consists of taking the sum of the ranks of the observations from a particular set  $x_1, x_2, \dots, x_n$  after they have been combined with the set  $y_1, y_2, \dots, y_n$  and the combination is arranged in ascending order. "T" may be calculated from the formula

$$T = \sum_{k=1}^n S_k - n\bar{v}$$

where  $S_k$  is the rank of  $x_k$  after the combination of the set of x's with the set of y's, and  $\bar{v}$  is the mean rank of the x's after combination. The probability of occurrence of any given rank total in the case of unpaired experiments with ranks 1, 2, --- n is found by taking all the possible totals beginning with the sum of the series 1, 2, --- n/2, and continuing by steps of one

up to the highest possible value  $(\frac{3}{4}n^2 + \frac{n}{2})/2$  and computing the numbers of ways in which each total may be obtained. If  $g_i$  is the number of ways that the  $i$ th total can be obtained and  $T$  is the total number of ways that all the  $\left\{ \left[ \left( \frac{3}{4}n^2 + \frac{n}{2} \right) / 2 \right] - \left[ \left( \frac{3}{4}n^2 + \frac{n}{2} \right) / 2 \right] \right\} = n^2/2$  totals can occur  $\cdot \sum_{i=1}^{n^2/2} g_i$ , the probability of occurrence of the  $i$ th total equals

$$P(t_i) = \frac{g_i}{\sum_{i=1}^{n^2/2} g_i} = \frac{g_i}{T}$$

Wilcoxon also shows that the probability of occurrence of any total or lesser total by chance under the assumption that there is no difference in means is given by the formula

$$P = 2 \left\{ 1 + \sum_{j=1}^r \sum_{i=1}^q \pi_j^i - \sum_{x=1}^{q-1} ([r-q-x+1] \pi_{q-x}^{r-x+x}) \right\} / \frac{(2q)!}{q!q!}$$

where  $\pi_j^i$  represents the number of  $j$ -part partitions of  $i$

$r$  is the serial number of possible rank totals 1, 2, ---  $r$

$q$  is the number of replicates

$n$  is an integer representing the serial number of the term in the series.

This probability may be represented more compactly by using the previous notation, as follows:

$$P(t_1 + t_2 + \dots + t_i) = \frac{\sum_{j=1}^i g_j}{\sum_{j=1}^{n^2/2} g_j}$$

Wilcoxon has tabulated some of these probabilities in his paper cited in the bibliography.

The Mann-Whitney U test, which for equal numbers of observations from each population being investigated is essentially the same as the Wilcoxon test, is defined by considering the quantities  $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$  arranged in order.

U is defined as the number of times a y precedes an x. The relationship between the Mann-Whitney test and the Wilcoxon test can be expressed, as follows

$$U = mn + \frac{n(n+1)}{2} - T$$

where T is the value given by the Wilcoxon test. The Mann-Whitney test is not, however, restricted to cases in which  $n = m$  as is the Wilcoxon. U has been shown to be a consistent statistic since the P-limit of the probability of rejecting the null-hypothesis when the null-hypothesis is false is 1 as  $n$  and  $m$  approach infinity.

## II. Purpose

The purpose of this thesis is the determination of the efficiency of a particular small sample of observations and to gain therefrom an insight into the approximate efficiency of these rank order tests for small samples when the population means of the two groups differ by fixed amounts. This problem will be approached experimentally because all attempts to use mathematics in the solution have thus far been held at an impasse because of the difficulties encountered in the integration of portions of the expression which result when the various considerations and relationships involved in the problem are expressed in mathematical terms.

In the analysis of the results of this experiment, the 10% level will be considered as the level of significance in the computation of the efficiency since the sample size selected does not lend itself to the use of the 5% level of significance. The reader may possibly wish to consider half of the efficiency at the 10% level of significance as roughly indicative of what the efficiency at the 5% level would be. The use of the 10% level of significance in non-parametric tests does not seem unrealistic, because, in general, non-parametric statistics tend to be more conservative than parametric statistics. In case non-parametric methods are applied to samples from a population which is normally distributed, the conservativistic tendencies of non-parametric inference validate the use of the higher significance

level.

The results of this experiment should prove helpful in the setting up of non-parametric quality control systems for such activities as subjective testing between a control group and a group receiving some particular treatment, the effects of which cannot be measured in any definite units.

In addition to information of a general nature concerning these tests, this thesis also hopes to provide relatively concrete information about the particular test in which three observations are taken from each group. Some insight into the efficiency of these tests for this particular sample size is of interest because in industrial applications of non-parametric quality control limits, it is frequently possible to fix the sample size.

### III. Method

The particular group of combinations which will be investigated in this thesis are those combinations which occur when three samples are drawn at random from each of two approximately normal populations. These samples will be assembled according to their rank and the particular test value assigned to the resulting combinations will be recorded. It may be readily seen that the number of possible outcomes when this procedure is followed is the number of combinations that can be made up of six things taken three at a time or --

$$N_c = C_3^6 = \frac{6!}{3!3!} = 20$$

different ways. These various combinations are given in figure (4) on page (18) where the letters R and W refer to observations drawn from a population R consisting of red balls and from a population W consisting of white balls, respectively. In this experiment, the samples will be drawn from a population which is approximately normal. We will make five sets of drawings in which each set will consist of a thousand observations. In each of these sets of drawings, the difference between the means of the populations from which the samples will be drawn will be assigned a given value (0, 0.5, 1.0, 1.5, 2.0 standard deviations). The frequency with which particular combinations occur under these differences in means will be observed and recorded. Special interest will be accorded to the particular combinations which are significant since they are a measure of the efficiency

of the tests.

The approximately normal populations from which the samples are to be drawn will consist of small balls which will be numbered in such a way that the frequency with which a particular number will appear will be approximately normal. Each of these populations will consist of nine hundred and ninety-six balls. The populations will be truncated at  $\pm 2.7$  standard deviations from the mean arbitrarily.

The following method will be used to obtain a truncated finite normal distribution from which the samples will be drawn. There will be fifty-four class intervals, from  $+2.7$  standard deviations to  $-2.7$  standard deviations from the mean. Each class interval will represent 0.1 standard deviations along the abscissa. There will, of course, be no class interval to correspond with the integer 0. The point between the  $-1$ st interval and the  $+1$ st interval will be considered as the mean. The number of balls that will be assigned to a particular class interval will be proportional to the ratio of the area under the normal curve above that particular interval on the abscissa, to the total area under the normal curve between the points  $\pm 2.7$  standard deviations on the abscissa. This criterion for determination of the number of balls per class interval may be expressed mathematically by the expression:

$$(a) \quad N_i = \frac{\int_{t_{i-1}}^{t_i} \phi(t) dt}{\int_{-2.7}^{+2.7} \phi(t) dt} \cdot N \quad -26 < i < 27$$

where  $N_i$  is the number of balls in the  $i$ th class interval when

$i > 0$  and the number of balls in the  $(i+1)$ st class interval when  $i \leq 0$  and where  $\phi(t)dt$  is the integrand of that function of  $t$ ,  $\phi(t)$ , which gives the normal curve of the type:

$$\phi(t) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{t^2}{2\sigma^2}}$$

and where  $N$  is the number of balls which in the case of this thesis was 1000 before the rounding off of the values shown in Figure (1). Since the area under a distribution function may be considered as being synonymous with probability, we may express the above formula in terms of probabilities, as follows:

$$N_i = \frac{1000 P_i}{\sum P_i} \quad -27 \leq i \leq 0 \leq i \leq 27$$

where  $P_i$  is the probability of occurrence of an observation within the limits of the  $i$ th class interval. For continuous distributions  $P_i$  is seen to be equal to  $\int_{i-1}^i \phi(t)dt$ , for positive values of  $i$ .

The values for these integrals have been tabulated by Kenney in his "Mathematics of Statistics". (5) The calculations involved in the construction of the population are given in tabular form by Figure (1). Since the normal distribution is symmetric, only the values for the positive half are given; the values for the negative half are the same. For purposes of calculation, the formula (a) was modified, as follows:

$$N_i = \frac{\int_0^i \phi(t)dt - \int_0^{i-1} \phi(t)dt}{\int_{-27}^{27} \phi(t)dt} \times 1000$$

Figure (2) shows in histogram form the population from which the samples will be drawn. The closeness of this popu-

Class Mark(i)	$\int_0^i \phi(t) dt$	$\int_0^{i-1} \phi(t) dt$	$\int_{i-1}^i \phi(t) dt$	Ni
1	.03983	.00000	.03983	40
2	.07926	.03983	.03943	40
3	.11791	.07926	.03865	39
4	.15542	.11791	.03751	38
5	.19146	.15542	.03604	36
6	.22575	.19146	.03429	34
7	.25804	.22575	.03229	32
8	.28814	.25804	.03010	30
9	.31594	.28814	.02760	28
10	.34134	.31594	.02540	25
11	.36433	.34134	.02399	24
12	.38493	.36433	.02060	20
13	.40320	.38493	.01827	18
14	.41924	.40320	.01604	16
15	.43319	.41924	.01395	14
16	.44520	.43319	.01201	12
17	.45543	.44520	.01023	10
18	.46407	.45543	.00864	9
19	.47128	.46407	.00721	7
20	.47725	.47128	.00597	6
21	.48214	.47725	.00489	5
22	.48610	.48214	.00396	4
23	.48928	.48610	.00318	3
24	.49180	.48928	.00253	3
25	.49379	.49180	.00199	2
26	.49534	.49379	.00155	2
27	.49653	.49534	.00119	1

HISTOGRAM  
OF  
APPROXIMATION  
TO  
NORMAL CURVE

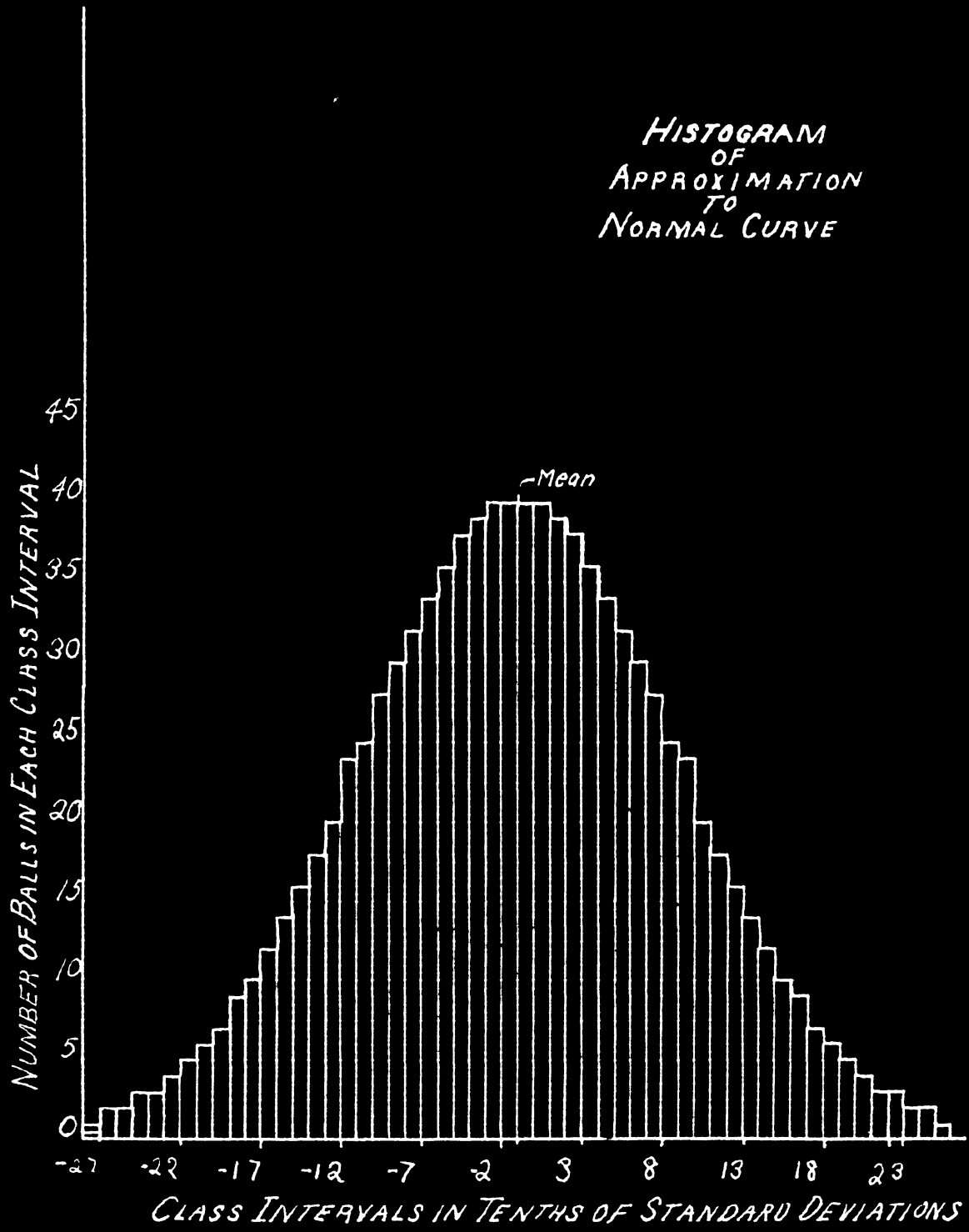


Fig 2

lation to a normal distribution can be readily seen.

The required number of balls will be marked for each class interval on small balls for two populations. Samples were drawn by means of a specially constructed paddle, three from each population. In the case of the null-distribution ( $M_1 = M_2$ ) these samples are arranged in descending order. This arrangement in descending order inverts the test number pattern for the Wilcoxon and Mann-Whitney tests. In the cases in which there is to be a difference in means, the desired difference is added to each of the observations taken from one of the populations and then the balls are arranged in descending order and the particular combinations which result are observed. It will be observed that for the purposes of this test we might just as well subtract the desired difference from the observations of one of the population, since the result would be an increase in the number of combinations in which the observations from the other population ranked lower than those of the transformed population, with the result that a different, but equally significant, test value would result, and the observed efficiency would be the same. It will also be observed that the particular distribution of the frequencies of occurrence of the various combinations that result under the null-distribution will be rectangular. When the means are different, the frequencies of combinations tends toward those particular groups which will indicate the lack of homogeneity in the underlying populations. This would, of course, introduce some slight error since the values of one population will be extended beyond

the point of truncation of the other population; however, this error will be inconsequential when compared with the error introduced in rounding off the values of the areas under the probability curve for each class interval to three decimal places for an approximation to a normal population. This will be seen, if the reader will take the sum of the squares of the area under the normal curve for those class intervals for which overlapping has occurred and compare it with the product of the absolute difference of the rounded-off value from its real value as given by tables for all the class intervals, times the probability of the occurrence of the particular class interval.

For the case of  $(M_1 - M_2) = 2.00$ , this is shown in tabular form by figure (3). Since the case where  $(M_1 - M_2) = 2.00$  is the one in which the overlapping error will be the largest, it is seen that this error is not of importance in this experiment.

For the purposes of this thesis, the definition of efficiency as a property of a statistic is given by considering two particular statistics G and K, which are estimators of the same parameter. The statistic G will be said to be more efficient than the statistic K if the probability of obtaining a significant value of G when the null-hypothesis is not true is greater than the probability of getting a significant value for the statistic K. Relative efficiency is defined as the ratio of the efficiency of one statistic to another when both estimate the same parameter. A statistic most frequently used as the standard of efficiency is  $t = \sqrt{N}(x - \bar{x})/\sigma$ .

Class Mark (i)	$P_i \left( \int_1^i - \frac{1}{\rho(t)} dt - N_i \right)$	Class Mark (i)	$(P_i)^2$
1	0.67711 X 10 <sup>-5</sup>	28	0.08281 X 10 <sup>-5</sup>
2	2.24751 X 10 <sup>-5</sup>	29	0.04761 X 10 <sup>-5</sup>
3	1.35275	30	0.02704
4	1.83789	31	0.01444
5	0.14416	32	0.00784
6	0.99441	33	0.00441
7	0.93641	34	0.00196
8	0.30100	35	0.00121
9	0.55600	36	0.00049
10	1.01600	37	0.00025
11	0.02399	38	0.00016
12	1.23600	39	0.00004
13	0.49329	40	0.00004
14	0.06416	41	0.00001
15	0.06975	42	0.00001
16	0.01201	43	0.00001
17	0.23529	44	0.00000
18	0.31104	45	0.00000
19	0.15141	46	0.00000
20	0.01791	47	0.00000
21	0.05379		
22	0.01584		
23	0.05724		
24	0.12096		
25	0.00199		
26	0.06975		
27	0.02261		
<hr/>		<hr/>	
	13.02037 X 10 <sup>-5</sup>		0.18933 X 10 <sup>-5</sup>

$$\text{Ratio} = \frac{0.18933}{13.02037}$$

Fig 3

#### IV. Results

The results obtained from this experiment are given here in the forms of tables and graphs. These results will be discussed in the following section. Figure (4) contains a recapitulation of the frequencies with which the various combinations occurred under particular differences in means. It also shows the test values which the different combinations assume under the Wald-Wolfowitz, Mann-Whitney, and Wilcoxon tests, respectively. Figures (5), (6), and (7) show the frequency with which the individual test values for the Wald-Wolfowitz, Mann-Whitney, and Wilcoxon tests have occurred under the various differences in means. It will be observed that the frequencies for the various test values in the Wilcoxon test, Figure (7), are the same as the frequencies for the corresponding test value in the Mann-Whitney test, Figure (6). For the remainder of this thesis, the Wald-Wolfowitz and Mann-Whitney tests only will be discussed, but the reader should remember that what is true of the Mann-Whitney test is also true of the Wilcoxon test when equal samples are taken from the different populations. When unequal samples are taken, the Wilcoxon test does not apply. In general, it is more easy to conduct a Mann-Whitney test on small numbers of observations than it is to conduct a Wilcoxon test. Figure (8) shows in histogram form the frequencies with which the various test values for the Wald-Wolfowitz test occur under the prescribed difference in means. Figure (9) is a composite of the five histograms shown in Figure (8). Figures (10) and (11) contain the

corresponding histograms for the Mann-Whitney test. The information and observations gained from these histograms will be discussed in detail in the succeeding section. Figure (12) shows the relationship between efficiency in percent and the various differences in means. For purposes of later discussion, the regression equation and the correlation of this relationship have been calculated, as follows:

y = % efficiency	X = (M <sub>1</sub> - M <sub>2</sub> )	y - $\bar{y}$	x - $\bar{x}$
13.7	0.5	-22.075	-0.75
28.3	1.0	- 7.475	-0.25
43.1	1.5	7.325	0.25
58.0	2.0	22.225	0.75

ba = g where

$$a \text{ is } \sum_j x_i x_j = \sum x_i^2$$

$$g \text{ is } \sum_j y_i x_j = \sum y_i x_i$$

b is the matrix of coefficients of x's

$$a = 1.25$$

$$g = 36.925$$

$$b(1.25) = 36.925$$

$$b = (36.925)/(1.25) = 29.54$$

$$y = 35.775 + 29.54(x - 1.25)$$

$$= 29.54 x - 36.925$$

$$y = 29.54 x - 1.15$$

An analysis of variance of this regression equation is shown in the succeeding section. The regression of (M<sub>1</sub> - M<sub>2</sub>) on percent efficiency is given by the calculations:

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$$

$$S_{xy} = 215.80$$

$$S_x^2 = 7.50$$

$$S_y^2 = 6210.19$$

$$\sqrt{S_x^2 S_y^2} = 215.82$$

$$r = \frac{215.80}{215.82} = .9999073$$

Table of Results I

COMBINATION	Difference in Mean					Values for tests		
	0	0.50	1.00	1.50	2.00	W-W	M-W	W
R R P W W W <sup>x1</sup>	52	131	279	431	580	2	0	6
R R R W R W W	58	68	150	165	192	4	1	7
R R R W W R W	50	91	110	84	51	4	2	8
R R R W W W R	48	57	52	28	10	3	3	9
R R W R W W R	45	28	12	6	2	5	4	10
R R W W R W R	48	57	35	11	1	5	5	11
R R W W W R R	50	40	16	6	8	3	6	12
R R W R R R W	51	91	88	114	65	4	2	8
R R W W R R W	43	85	40	28	9	4	4	10
R R W R R R W	44	80	69	28	18	6	3	9
R W W W R R R <sup>x1</sup>	48	6	4	0	0	2	9	15
W W R R R R R	47	16	5	2	0	4	8	14
W W R R R W R	31	19	7	0	0	4	7	13
W W R R R R W	59	28	30	5	1	3	6	12
W R R W R R W	56	34	6	11	1	5	5	11
W R R R W R W	55	28	5	5	13	5	4	10
W R R R R W W	43	62	75	40	38	3	3	9
W R R W W R R	54	23	0	7	0	4	7	13
W R R R W R R	51	40	17	23	11	4	5	11
W R W R W R R	47	16	0	6	0	6	6	12

Fig. 4 Table of results: Frequency with which the various combinations were drawn when the two population means differed by the given amount.

<sup>x</sup> significant sets

<sup>x1</sup> Significant sets under the hypothesis that  $M_R$  is greater than  $M_W$

## Table of Results II

## Wald-Wolfowitz Test:

Test Number	0	Difference in Mean			
		0.50	1.00	1.50	2.00
2 <sup>x</sup>	100	137	283	431	580
3	200	187	173	84	57
4	405	433	399	423	328
5	204	147	58	33	17
6	91	96	69	34	18
	1000	1000	1000	1000	1000

Fig. 5 Wald-Wolfowitz test on results shown in Fig.

## Mann-Whitney Test:

Test Number	0	Difference in Mean			
		0.50	1.00	1.50	2.00
0 <sup>x</sup>	52	131	279	431	580
1	58	68	150	165	192
2	101	182	198	198	116
3	135	199	186	96	66
4	143	141	57	39	24
5	155	131	58	45	13
6	156	84	46	17	9
7	105	42	7	7	0
8	47	16	5	2	0
9 <sup>x</sup>	48	6	4	0	0

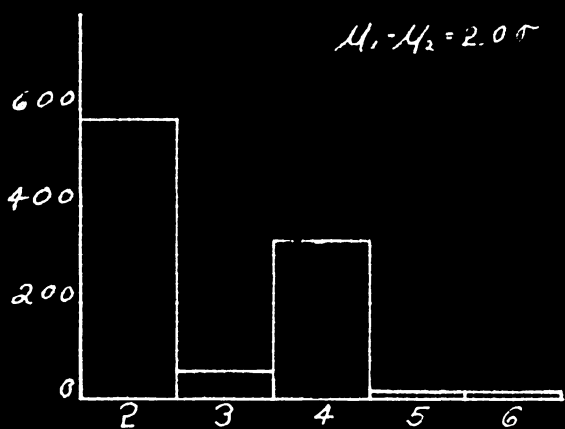
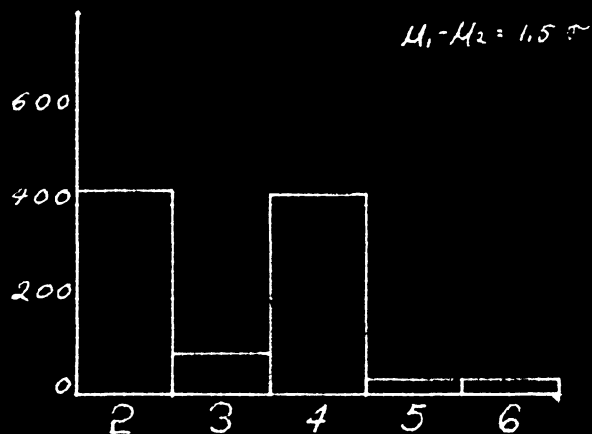
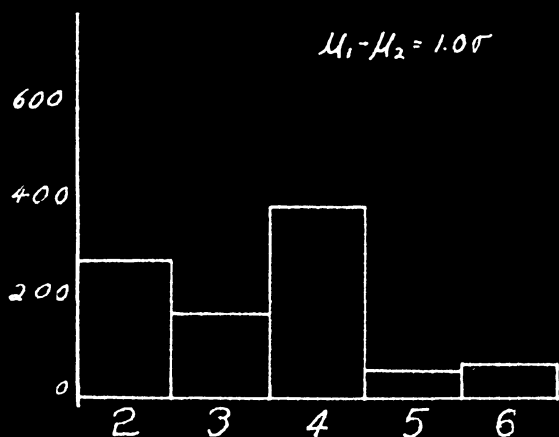
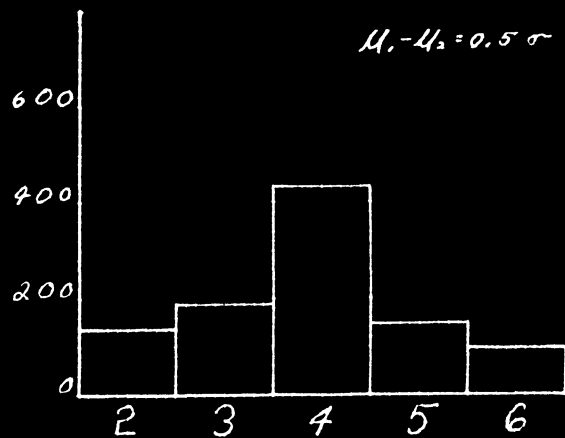
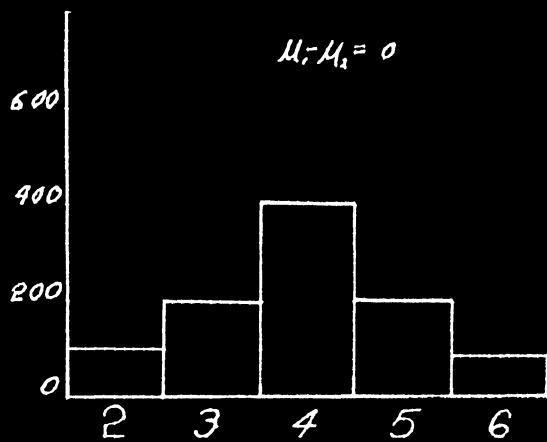
Fig. 6 Mann-Whitney test on results shown in Fig.

## Wilcoxon Test:

Test Number	0	Difference in Mean			
		0.50	1.00	1.50	2.00
6 <sup>x</sup>	52	131	279	431	580
7	58	68	150	165	192
8	101	182	198	198	116
9	135	199	186	96	66
10	143	141	57	39	24
11	155	131	58	45	13
12	156	84	46	17	9
13	105	42	7	7	0
14 <sup>x</sup>	47	16	5	2	0
15 <sup>x</sup>	48	6	4	0	0

Fig. 7 Wilcoxon test on results shown in Fig.

x significant groups



WALD-WOLFOWITZ  
TEST VALUES FOR VARIOUS  
( $\mu_1 - \mu_2$ )  
FREQUENCY VS. VALUE

Fig 8

HISTOGRAM SHOWING  
 FREQUENCY OF VALUES IN THE  
 WALD-WOLFAWITZ TEST  
 FOR DIFFERENT  
 $(\mu_1 - \mu_2)$

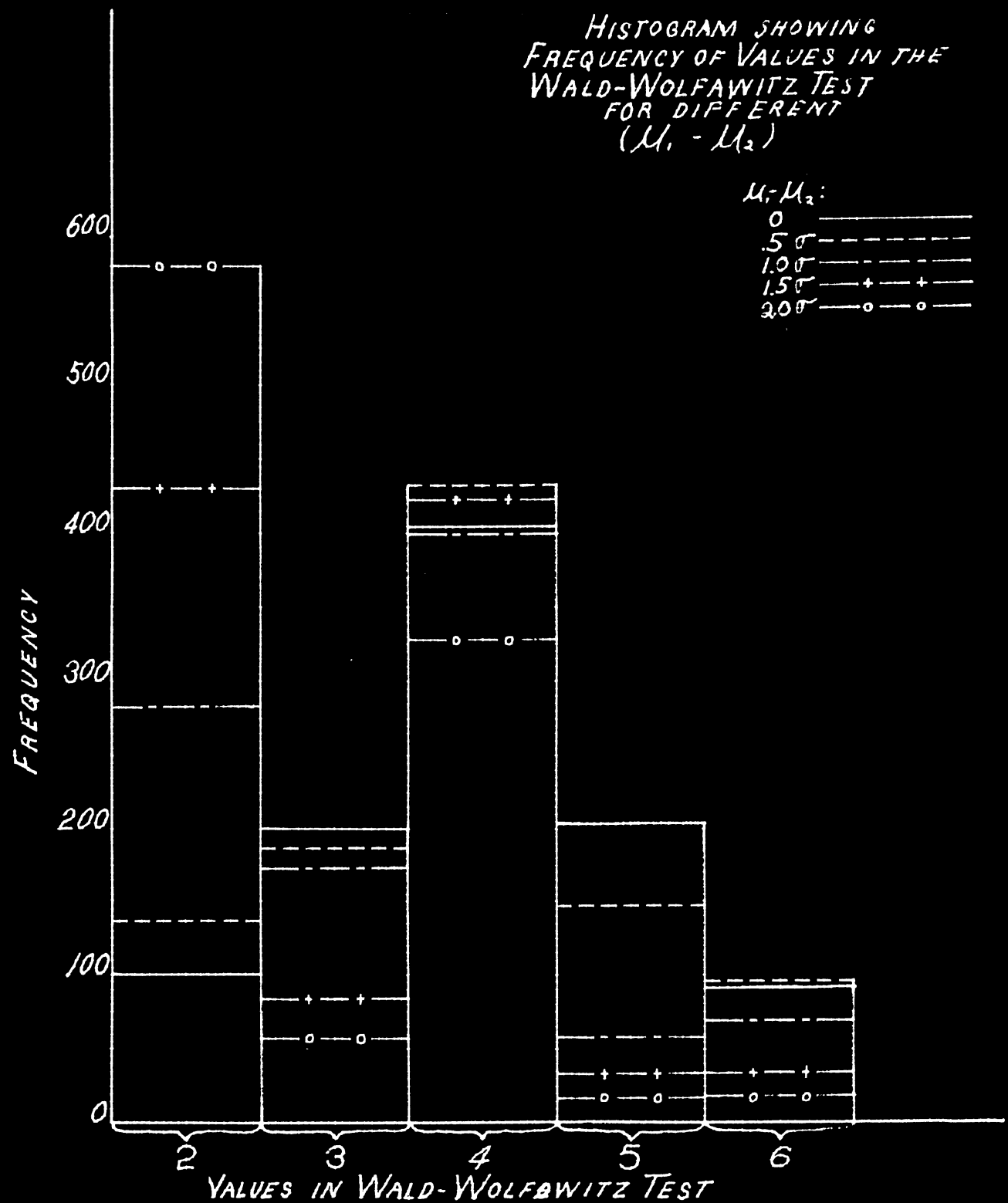
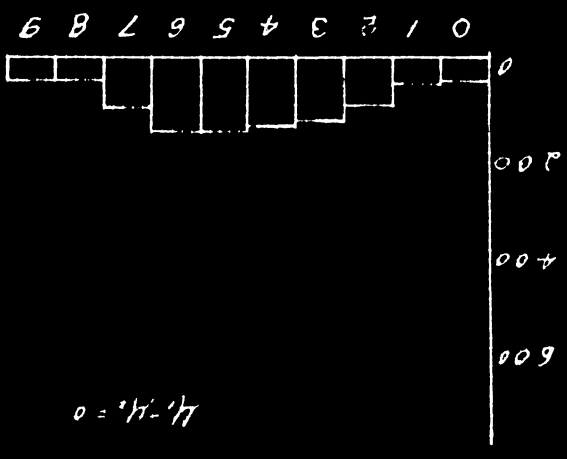
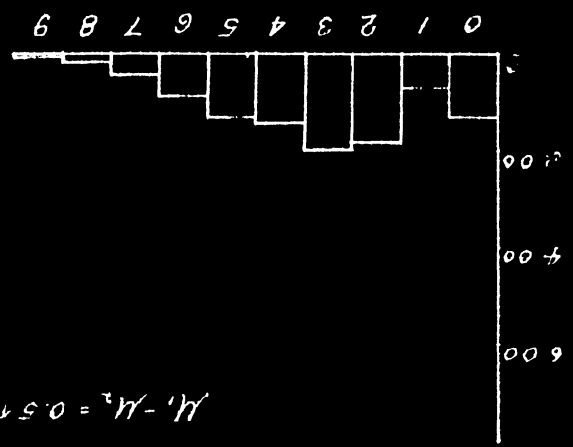
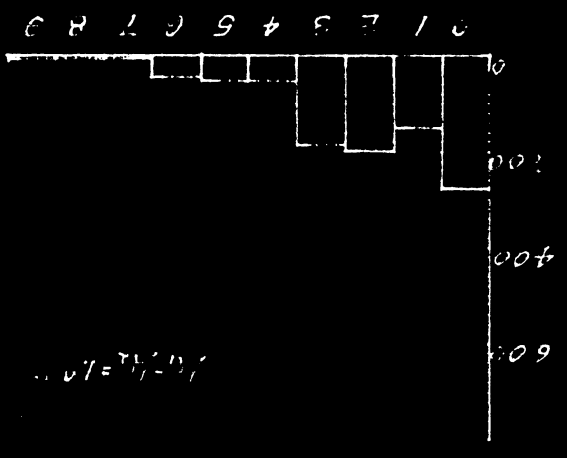
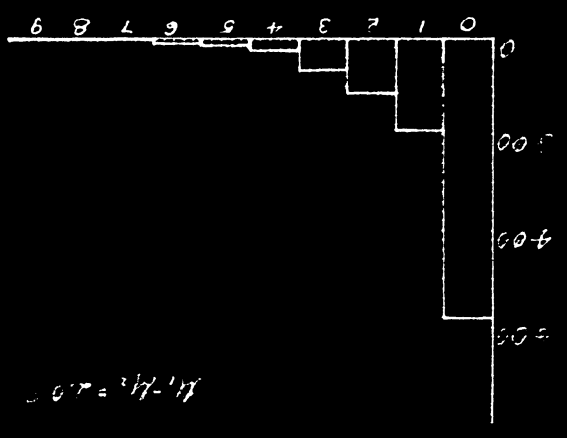
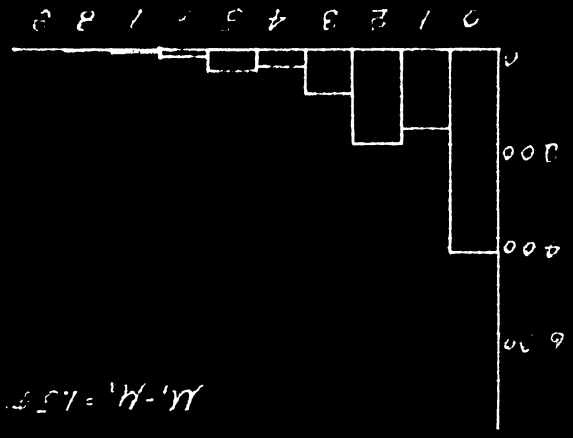


Fig 9

Fig 10

MANN-WHITNEY  
TEST VALUES FOR VARIOUS  
 $M_1 - M_2$   
FREQUENCY VS. VALUE



HISTOGRAM SHOWING  
 FREQUENCY OF VALUES IN THE  
 MANN-WHITNEY TEST  
 FOR DIFFERENT  
 $(\mu_1 - \mu_2)$

$\mu_1 - \mu_2$	Symbol
0	—
$0.5\sigma$	---
$1.0\sigma$	- - -
$1.5\sigma$	- + -
$2.0\sigma$	- o -

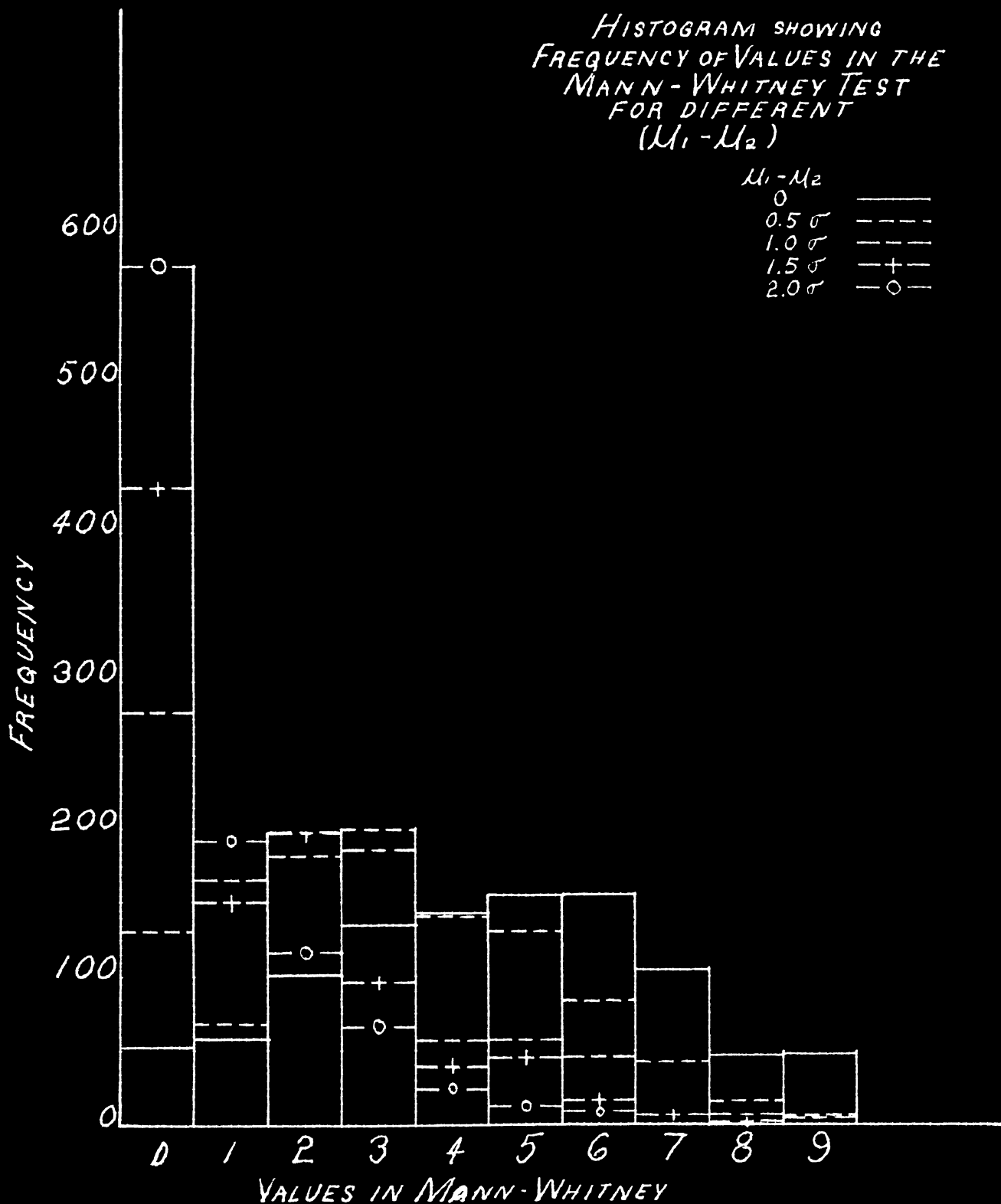


Fig 11

CURVE OF  
EFFICIENCY  
VS  
DIFFERENCE IN MEANS

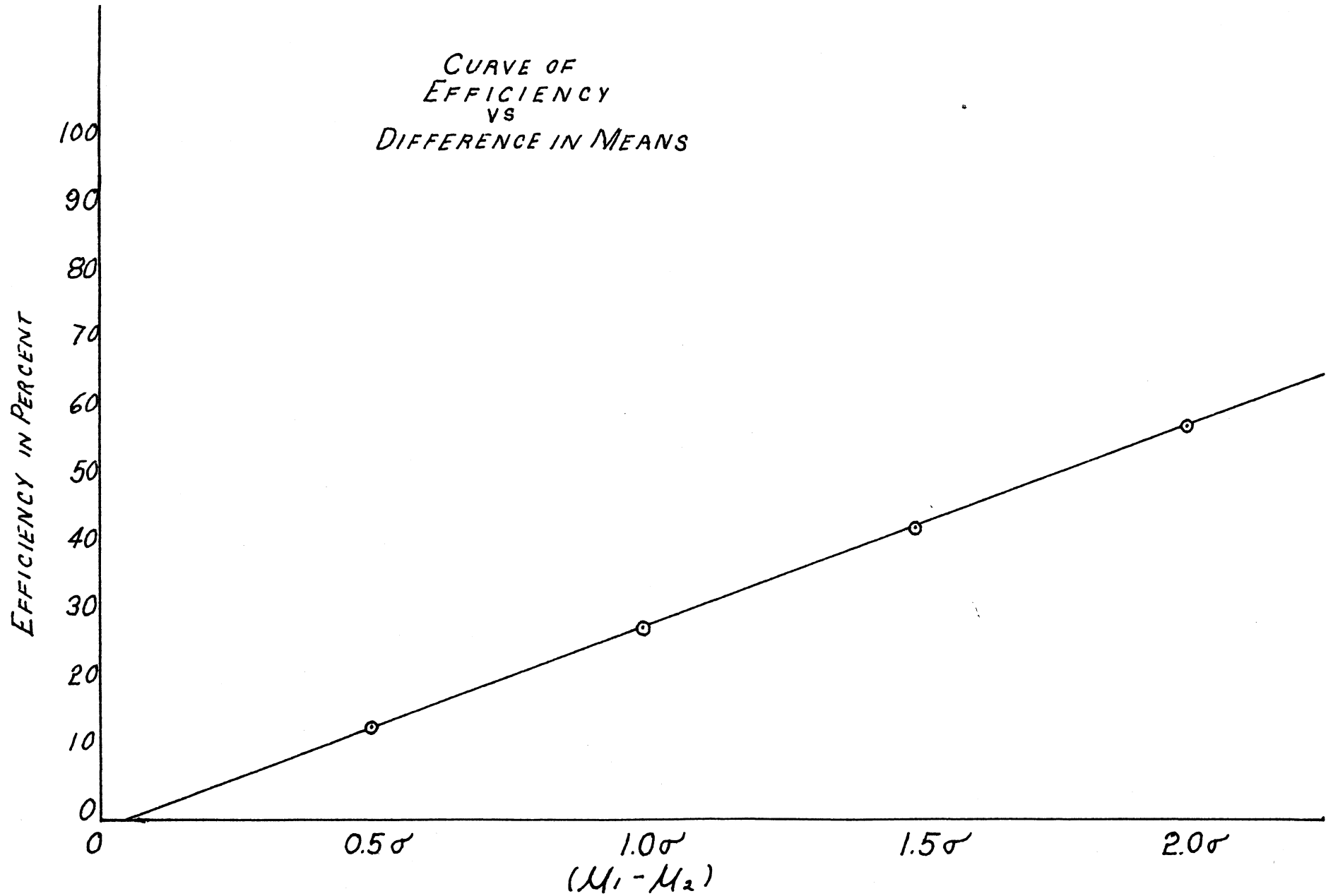


Fig 12

## V. Discussion of Results

Perhaps the most interesting result to be obtained from this experiment is extreme closeness to linearity of effect of an increase in the difference in means on the efficiency of the rank order statistics under surveillance. An analysis of variance on this regression shows:

Source	SS	D/F	Mean Square
Linear effect	1090.77	1	1090.77
Error	0.02	2	0.01
Total	1090.79	3	

$F = 109,077$  (significant)

This high degree of linearity for this particular set of observations should not be considered as a characteristic of all sets. These results are a function of too many variables to be assumed to represent a universal trend in efficiency vs. difference in mean. This result was a function of the number of balls selected from each population, the number of observations, the accuracy of the approximation to normality in the constructed population, etc. However, since in many cases the experimenter is able to select the number of observations to be made from each population, this high linearity of results should give him confidence regarding any interpolation on the curve when an estimation of efficiency for various differences in means are desired. This curve will probably not prove too useful for use in extrapolation as it is to be expected that the efficiency curve will flatten off soon after the 2.00 point has been reached.

An observation of the histograms which show the results of this experiment shows that the efficiency of these rank order statistics when the means of the two populations differ by a fixed amount are:

$M_1 - M_2$	Eff
0.50	13.7
1.00	28.3
1.50	43.1
2.00	58.0

The linearity of these results has already been noted in detail. A least squares equation on these results has been calculated and is:

$$Y = 29.54 X - 1.15$$

where Y is the efficiency and X is the difference in means of the two populations.

It will be observed that for the particular sample size chosen, the efficiencies of all the tests considered are the same. The selection of the most desirable tests must then be based on the criterion of ease in calculation. It appears that the Wald-Wolfowitz test will be best in this respect. However, under the hypothesis that the mean of one population is greater than that of the other, the Mann-Whitney test gives results that are significant at the 5% level.

VI. BIBLIOGRAPHY

1. Hotelling H.; (Text Unnamed), to be published at later date.
2. Wald, A. & Wolfowitz, U.; "On a Test Whether Two Samples are From the Same Population", P 147, Annals of Mathematic Statistics, Vol XI, 1940.
3. Wilcoxon, F.; "Individual Comparisons by Ranking Methods", P 80, Biometrics, Vol 1, No. 6.
4. Mann, H. & Whitney D.; "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other", P 50, Annals of Mathematic Statistics, Vol XVIII, No. 1, 1947.
5. Kenney, ; "Mathematics of Statistics", Part One, PP 235-7, D Van Nostrand, New York.