EXPERIMENTAL COMPARISON OF STUDENT's AND WILCOXON's TWO SAMPLE TESTS

J. HEMELRIJK

TECHNOLOGICAL UNIVERSITY, DELFT, NETHERLANDS

REPORT SP 74 OF
THE STATISTICS DEPARTMENT OF THE MATHEMATICAL CENTRE, AMSTERDAM

REPRINT FROM
QUANTITATIVE METHODS IN PHARMACOLOGY, NORTH HOLLAND PUBLISHING CO,
AMSTERDAM 1961

EXPERIMENTAL COMPARISON OF STUDENT's AND WILCOXON's TWO SAMPLE TESTS [1])

J. HEMELRIJK

Technological University, Delft, Netherlands

1. Introduction

In statistics, just like in the industry of consumer goods, there are producers and consumers. The goods are statistical methods. These come in various kinds and "brands" and in great and often confusing variety. For the consumer, the applier of statistical methods, the choice between alternative methods is often difficult and too often depends on personal and irrational factors.

The advice of producers cannot always be trusted implicitly. They are apt — as is only natural — to praise their own wares. The advice of consumers — based on experience and personal impressions — cannot be trusted either. It is well known among applied statisticians that in many fields of applied science, e.g. in industry, experience, especially "experience of a lifetime", compares unfavorably with objective scientific research: tradition and aversion from innovations are usually strong impediments for the introduction of new methods, even if these are better than the old ones. This also holds for statistics. Admittedly, the introduction of new methods necessitates investments: the "purchase" of a new statistical method costs the consumer time and work. The decision about these investments should, however, be based on the results to be expected and not on subjective considerations.

An objective appraisal of new methods as compared to the methods in general use can only be made if their practical properties are known to a sufficient degree. In this respect the situation in statistics is in certain cases unsatisfactory. Take for instance the problem of two samples. Many statistical tests have been developed for the hypothesis that two samples originate from the same population and although for most of these tests asymptotical properties concerning consistency and

[1]) Report SP 74 of the Statistics Department of The Mathematical Centre.

efficiency have been derived, the small sample properties — which are of far more practical importance — are known for a small number only. Thus the producers have not yet succeeded in providing the consumers with sufficient data to make a rational choice between existing alternative methods.

What would be needed would be something like a "consumers guide" listing the relevant properties of rival methods. Much relevant information can be found in textbooks and in papers scattered throughout statistical litterature. The compilation of a "consumers guide" from this material would not only be very useful for the consumers but it would also clearly reveal those points where sufficient knowledge is still lacking.

One such point is certainly to be found in the small sample properties of a comparatively new "brand" on the statistical market: non-parametric methods. Three of todays papers report on investigations of the properties of an important non-parametric test: WILCOXON's two sample test, and on comparisons of this test with the corresponding "normal" test of STUDENT. Although far from complete — research is still in progress — the amount of work invested in these investigations and their practical importance justify their presentation at this moment.

2. The Tests Compared

STUDENT's t-test does not need comment. It is universally known as the uniformly most powerful test for the comparison of the means of two independent samples drawn from normal distributions with equal variances (under "Student-conditions", for short). The sensitivity, under the null-hypothesis, to departures from normality and to inequality of the variances are in general judged not to be alarming as long as the sizes of the samples are approximately equal. The effect of these deviations on the power of the test has not been investigated deeply, but it is known that the property of being the most powerful test is lost.

WILCOXON's test, although widely known, may still need some explanation. The definition of the test statistic W is as follows. If $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ are the two samples, then W is equal to 2 times the number of pairs (x_i, y_j) with $x_i > y_j$ + the number of pairs

(x_i, y_j) with $x_i = y_j$. (Often $U = \frac{1}{2}W$ is used). The null-hypothesis is that the samples have been drawn from the same population; the parent distribution may be of any form, continuous or discrete. The test is, under the hypothesis tested, distributionfree.

It is clear that of the two Wilcoxon's test is the more general one and one should therefore not be surprised to find that in the special case of normality with equal variances its power is less than the power of Student's test, as the latter has been devised especially for those specific circumstances.

The asymptotic local efficiency of Wilcoxon's test as compared with Student's has been proved to be equal to $3/\pi = 0.95$ for the normal case; this means, roughly, that for large samples Wilcoxon's test needs about 5 % more observations than Student's test if the Student-conditions are satisfied and if one of the two distributions is only slightly shifted with respect to the other. This result and many more of the same kind can be found in G. E. NOETHER (1958), where it is also pointed out that under non-Student-conditions Wilcoxon's test may be far more efficient as a test against shift than Student's test.

The natural complement of these important asymptotical investigations are those for small samples. Mathematically, however, distributionfree methods, although generally simple under the hypothesis tested, are extremely complicated under alternative hypotheses. Theoretical results for small samples and alternative hypotheses are scarce and we have therefore taken recourse to a sampling experiment in order to compare the two tests for samples of size 10 each. Estimates were computed, under Student-conditions and for exponential distributions, of the power functions of the onesided tests with level of significance 0.025; of the number of times when the two tests lead to different results and of the sensitivity for slippage. This sampling experiment, which has been executed on ordinary desk calculators, is only a small one which we hope to extend in the future by means of more powerful equipment.

3. The Sampling Experiment

M. H. QUENOUILLE's (1959) "Tables of random observations from standard distributions" contain a sample of 1000 observations from a standard normal distribution and transformations of this sample into samples from seven other well known distributions. Corresponding

sample values for two different distributions provide the same percentile in each of these distributions. The standard normal distribution and the exponential distribution have been used for our sampling experiment. The latter distribution is very skew as may be seen from Fig. 1.

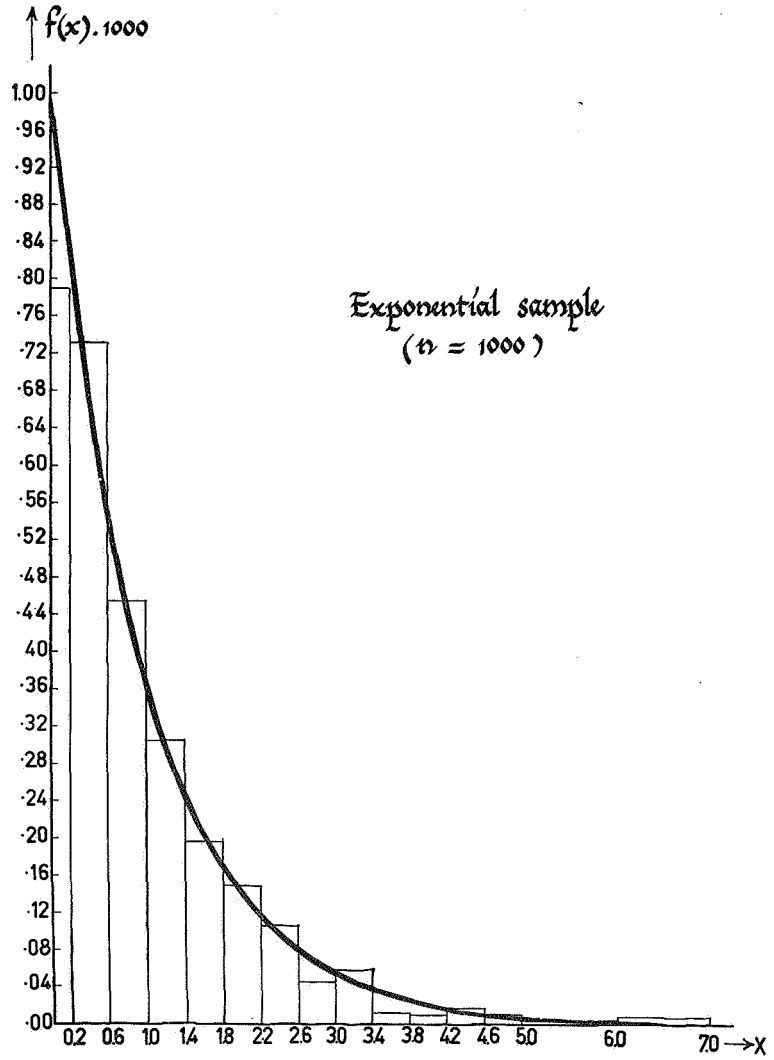


Fig. 1. Exponential distribution and sample of 1000 observations.

For both distributions the 1000 observations were split up into 50 pairs of samples of 10 observations each. In the order of Quenouille's table observations nr 1, . . ., 10 and 11, . . ., 20 formed the first pair of samples, etc. For each of these pairs of samples Student's t and Wilcoxon's W were computed, giving a pair of values (t, W). The value of W is the same for corresponding pairs of samples in the normal and in the exponential case, because W is a rank order statistic and the transformation from the normal to the exponential case is a monotonic transformation.

For the *normal case* the same computations were made after applying four different displacements δ to the second sample of each pair, in order to estimate and compare the power functions of the tests. The shifts chosen were

$$\delta = 0 \qquad 0.81 \qquad 1.05 \text{ and } 1.53.$$

These values were chosen such that the power of the relevant one-sided Student-test with level of significance $\alpha = 0.025$ assumed the values

$$0.025 \qquad 0.40 \qquad 0.60 \text{ and } 0.90 \text{ respectively.}$$

The original distribution and the shifted ones are shown in Fig. 2.

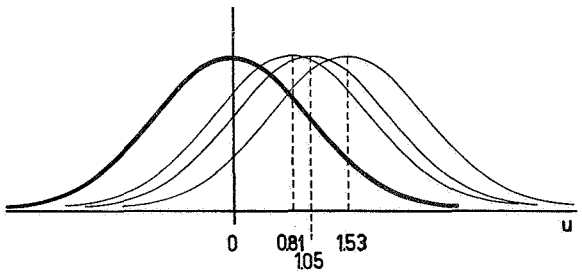
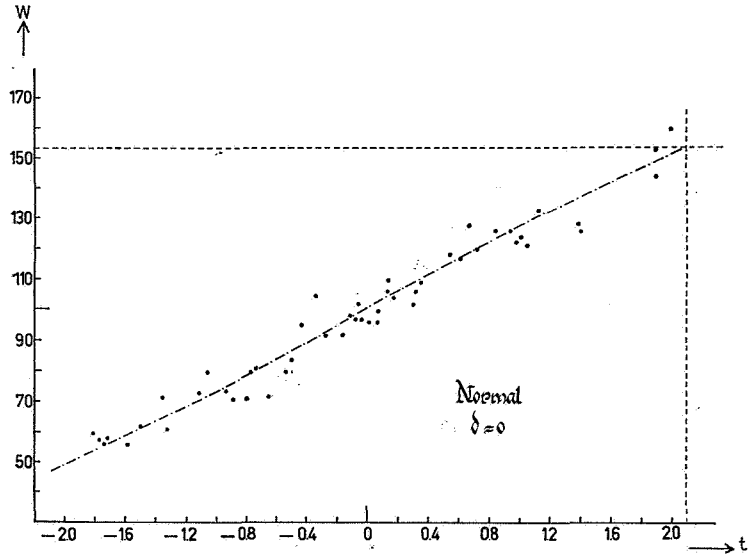
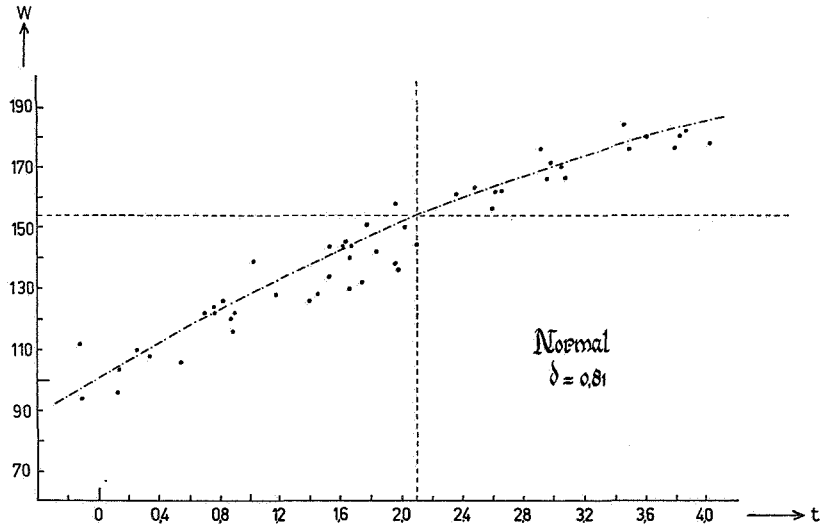


Fig. 2. Standard normal distribution and shifted distributions.

For each of these four situations the 50 pairs of values (t, W) found were plotted as shown in Fig. 3, 4, 5 and 6.

In each of these figures a curve has been drawn connecting pairs of critical values (t_r, W_r) pertaining to one-sided tests with the same level of significance for both tests. The expected values under the null hypothesis are 0 and 100 for t and W respectively. The critical values t_r and W_r for level of significance $\alpha = 0.025$ have been indicated by vertical and horizontal broken lines.


W
170
150
130
110
90
70
50
Normal
δ = 0
−2.0 −1.6 −1.2 −0.8 −0.4 0 0.4 0.8 1.2 1.6 2.0
t


W
190
170
150
130
110
90
70
Normal
δ = 0,81
0 0,4 0,8 1,2 1,6 2,0 2,4 2,8 3,2 3,6 4,0
t

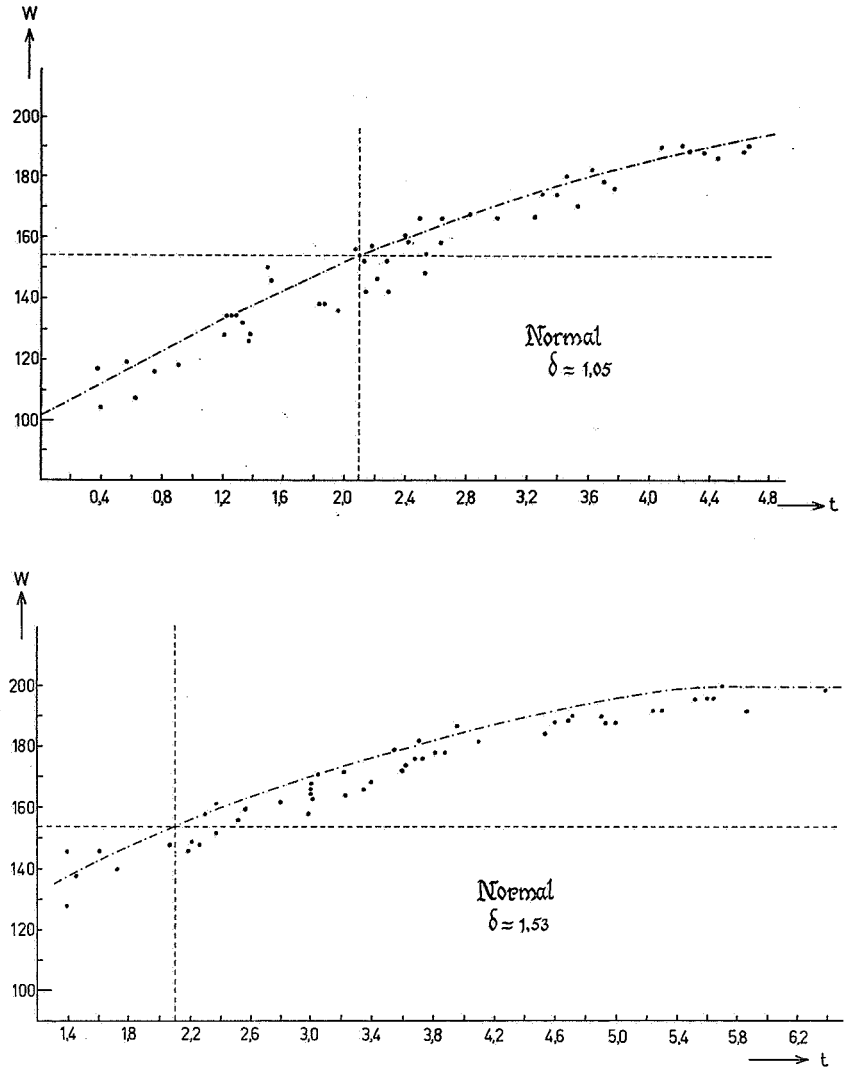


Fig. 3—6. Values of t and W for 50 pairs of samples of size 10 from standard normal distributions; displacement δ.

For $\delta = 0$ (the null-hypothesis) the points (t, W) cluster rather closely around the curve of equal levels of significance and the number of points above and below are approximately equal. For the other

three cases, with positive shift, the number of points below the curve becomes progressively larger, indicating a growing difference in power between the two tests. A point below the curve indicates that for this pair of samples the one-sided tail probability (or "P-value") is smaller for Student's test than for Wilcoxon's test, thus showing a greater power of the former. Points lying in the first quadrant formed by the broken lines indicate that both Student's and Wilcoxon's test (with $\alpha = 0.025$) lead to rejection of the null hypothesis; points in the third quadrant indicate rejection by neither, in the second rejection by Wilcoxon's but not by Student's test and in the fourth quadrant rejection by Student's but not by Wilcoxon's test.

Apart from these investigations concerning the power functions — which will be summarized later — the question of sensitivity for slippage was considered. To this end we returned to the original samples (with $\delta = 0$) and, for each pair, added 2 to the first random observation of the sample with the largest median. This sample was then considered as the second one of the pair and the same computations as above were executed. The results have been plotted in Fig. 7.

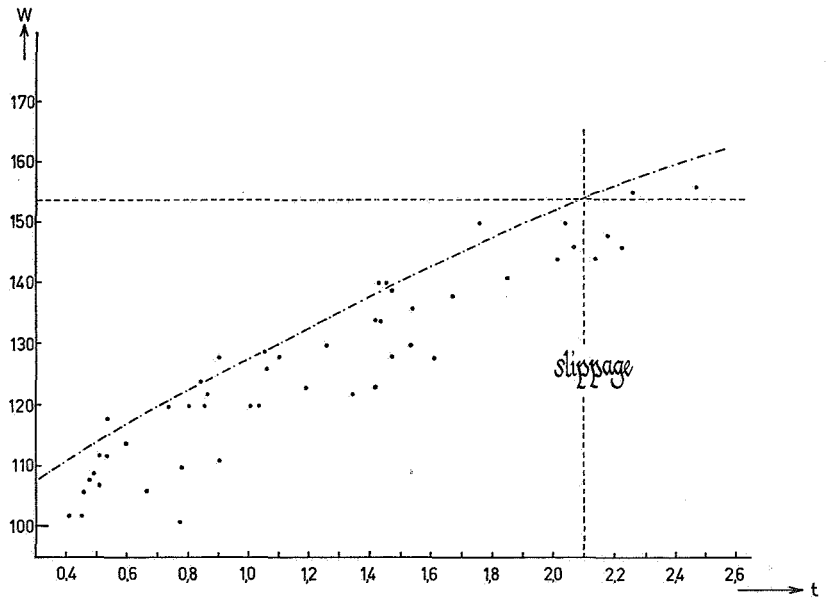


Fig. 7. Values of (t, W) after adding 2 to the first observation of the sample with the largest median.

It is evident from Fig. 7 that Student's test reacts far stronger to this treatment: most of the points are situated rather far under the line of equal levels of significance indicating a larger sensitivity for errors in the observations for Student's test.

Estimates of the values of the powerfunction (i.e. the probability of rejecting the hypothesis tested) can be derived from fig. 3, . . ., 7. These have been summarized in Table 1 and Fig. 8.

TABLE 1. *Theoretical and estimated power functions of* STUDENT's *and* WILCOXON's *two sample tests for* $m = n = 10$ *and* $\alpha = 0.025$; *normal case*

δ	shift (in terms of standard deviations) 0	0.81	1.05	1.53	slippage
Student	0 (0.025)	0.34 (0.40)	0.62 (0.60)	0.88 (0.90)	0.10
Wilcoxon	0.02 (0.022)	0.36	0.52	0.80	0.06

(Values between brackets are theoretical values)

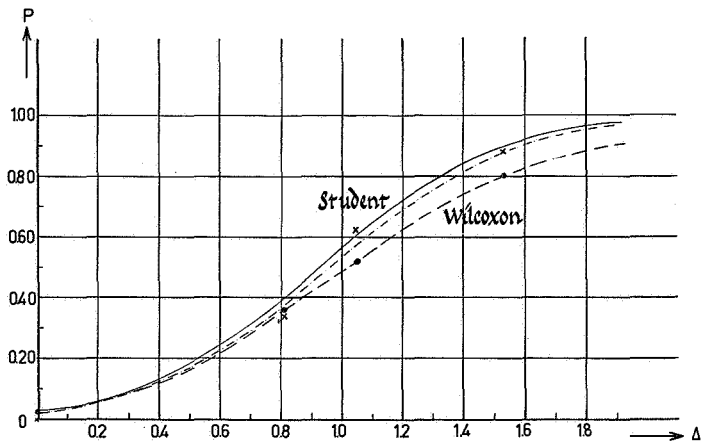


Fig. 8. Power functions ($m = n = 10$; $\alpha = 0.025$, one-sided tests).

In Fig. 8 the true power function of Student's test has been drawn. The estimated values (×) do not deviate too far from this curve.

A (broken) line has been drawn by hand through the three estimated points of the power of Wilcoxon's test. The results indicate that the power of the latter is in this case about 9/10 of the power of the former.

Owing to the discreteness of W the power under $\delta = 0$ (the "true level of significance") is 0.022, thus less than $\alpha = 0.025$, while for Student's t it is exactly equal to α. In order to see how much of the loss of power is due to this difference in size of the critical regions a third line (— · — ·) has been drawn: the power of Student's test with critical region of size 0.022. The situation remains essentially the same.

Table 1 also shows the stronger reaction of Student's test to an observational error. This aspect can be judged still better from Table 2.

TABLE 2. *One-sided tail probabilities; normal case*

δ	shift (in terms of standard deviations) 0	0.81	1.05	1.53	slippage
$S < W$	24	33	32	41	42
$W < S$	23	16	13	5	7
$S \approx W$	3	1	5	4	1

($S < W$ = tail probability of STUDENT's test smaller than of WILCOXON's, etc.)

As may be seen from this table the difference in sensitivity for an observational error of 2 standard deviations in one observation is in this respect comparable to the difference in power for $\delta = 1.53$. Comparison of Fig. 7 and 6 shows that the difference in sensitivity to slippage is the larger of the two.

In Table 3 the decisions, reached when testing with $\alpha = 0.025$, are summarized. In most cases the decisions according to Student's and Wilcoxon's test are the same, but especially for $\delta \geqq 1.05$ some are different, usually to the advantage of Student's test. In the case of slippage the advantage goes to Wilcoxon's test.

TABLE 3. *Decisions reached with $\alpha = 0.025$; normal case*

S	W	Shift (in terms of standard deviations) $\delta = 0$	0.81	1.05	1.53	slippage
—	—	49	31	18	6	44
+	+	0	17	25	40	2
same		49	48	43	46	46
+	—	0	1	6	4	3
—	+	1	1	1	0	1
different		1	2	7	4	4

Remark. From a personal communication of W. J. DIXON and D. TEICHROEW it appears that larger investigations of the same kind as described here (250 pairs of samples in total) led, for the case considered here, to somewhat different results. Whereas our data (cf. Fig. 4, 5 and 6) do not leave any doubt as to whether Student's test has, for the normal case, larger power than Wilcoxon's, DIXON and TEICHROEW do not find this difference. They eliminated the discrete character of W by means of a randomization procedure, in order to reach the exact level of significance 0.025, but this cannot fully explain the difference between their findings and our own. In Fig. 4, 5 and 6 such a procedure would result in a small shift of the curve of equal levels of significance, but most of the points (t, W) would remain below

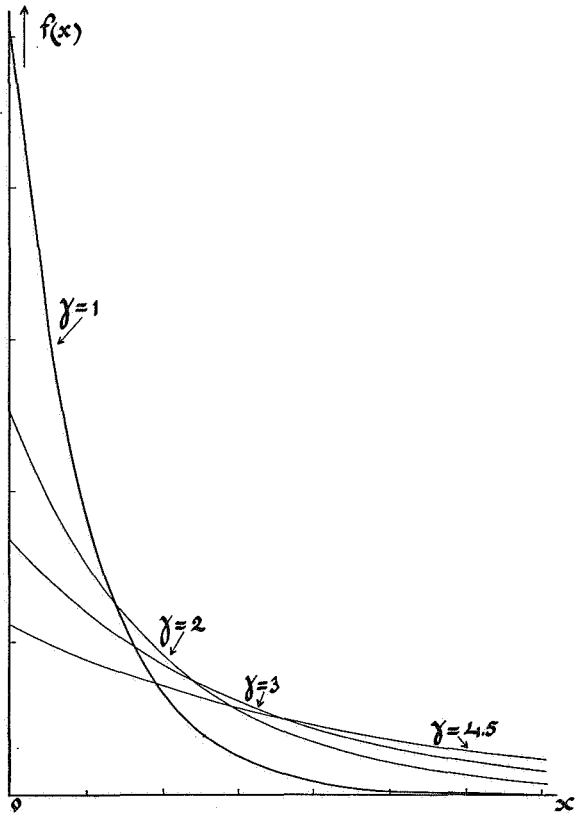


Fig. 9. Exponential distributions.

these lines. The discrepancy between their results (which also cover other distributionfree tests) and ours indicates that further research is necessary.

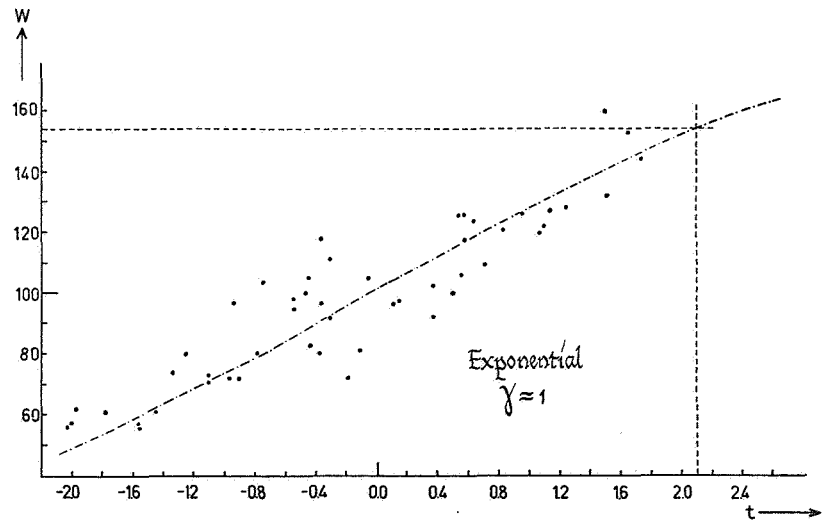
Similar computations as above were executed for *exponential parent distributions*. The starting point of this distribution being 0 the alternative hypotheses were not chosen in the form of shifts, but multiplication factors

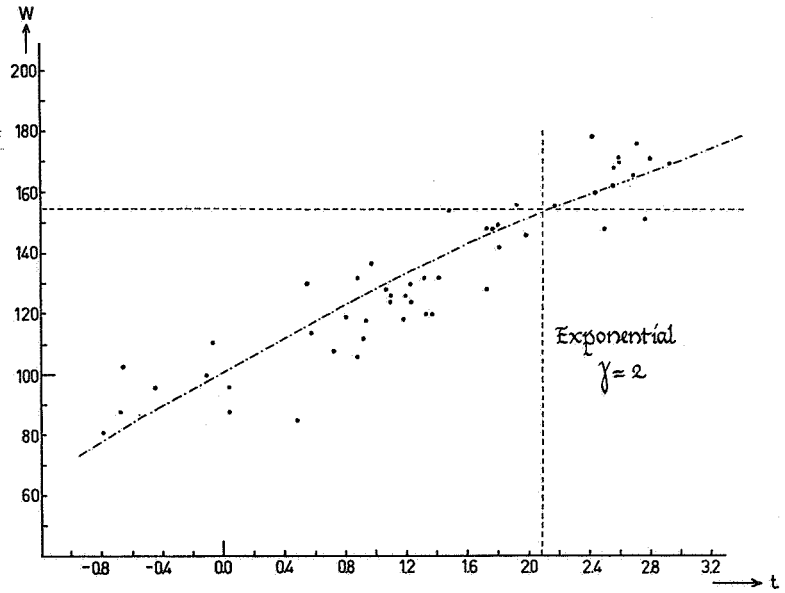
$$\gamma = 1 \quad 2 \quad 3 \text{ and } 4.5$$

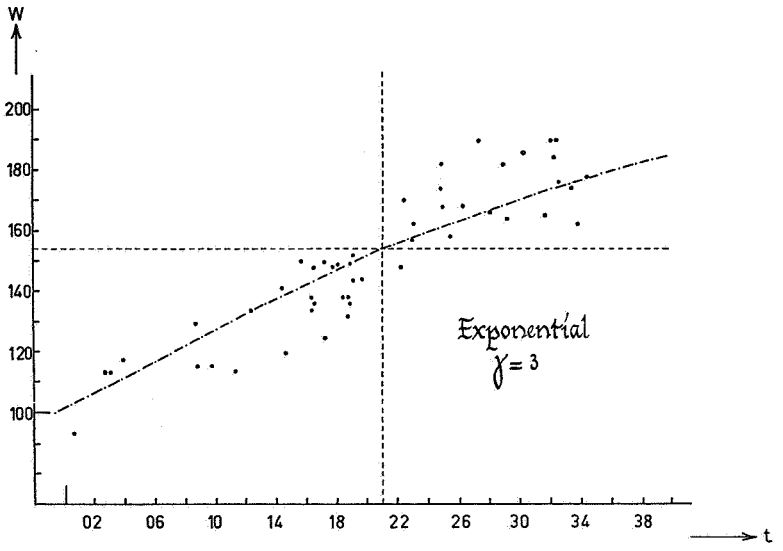
were applied to the second sample of each pair, as indicated in Fig. 9.

For the rest the procedure was the same as in the normal case except for the slippage, which was not repeated here. The results are summarized in Fig. 10, . . ., 13 and Table 4, 5 and 6.

Comparing these graphs with those for the normal case, the situation seems to be quite different. Especially in the region of small tail probabilities there are now, under the alternative hypotheses $\delta = 2$, 3 and 4.5, more points above the line of equal levels of significance than below.




W
200
180
160
140
120
100
80
60
-0.8
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
t
Exponential
γ = 2


W
200
180
160
140
120
100
02
06
10
14
18
22
26
30
34
38
t
Exponential
γ = 3

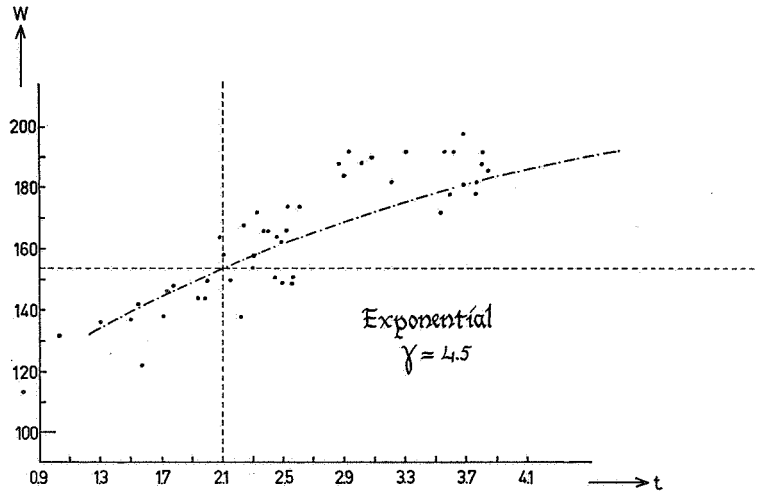


Fig. 10—13. Values of t and W for 50 pairs of samples of size 10 from exponential distributions (multiplication factor γ for the second sample).

TABLE 4. *Estimated power functions of* STUDENT'S *and* WILCOXON'S *two sample tests for* $m = n = 10$ *and* $\alpha = 0.025$; *exponential case*

γ	multiplication factor			
	1	2	3	4.5
Student	0	0.26	0.44	0.74
Wilcoxon	0.02	0.26	0.42	0.64

TABLE 5. *One-sided tail probabilities; exponential case*

γ	multiplication factor			
	1	2	3	4.5
$S < W$	23	25	23	20
$W < S$	25	22	24	29
$S \approx W$	2	3	3	1

The favourable result for $\gamma = 4.5$ for Student's test seems to be somewhat fortuitous in this case. For $\alpha = 0.005$ e.g. the advantage would have been on the side of Wilcoxon's test as may be seen from Fig. 13. Also Table 5, where no level of significance is specified, indicates a slight preponderance of Wilcoxon's test.

TABLE 6. *Decisions reached with* $\alpha = 0.025$; *exponential case*

S	W	multiplication factor $\gamma = 1$	2	3	4.5
−	−	49	35	28	12
+	+	0	11	21	31
same		49	46	49	43
+	−	0	2	1	6
−	+	1	2	0	1
different		1	4	1	7

Conclusions

It is clear that from so small an experiment no clearcut and final conclusions may be drawn. In general, however, our results agree with expectations based on asymptotic results obtained theoretically: Student's test is more powerful for small samples under Student-conditions, but the situation is somewhat reversed for the skew exponential distributions considered. Also the sensitivity, under the null-hypothesis, for an observational error, is larger for Student's test than for Wilcoxon's.

Even though these conclusions cannot be final — especially because they are not confirmed by the investigations of Dixon and Teichroew — they do contain an indication for further research. They show for instance that it may be profitable, in situations resembling Student-conditions, to use TERRY's or VAN DER WAERDEN's variations on Wilcoxon's theme, making the test asymptotically equivalent to Student's test (under Student-conditions), thus probably improving its power for small samples without losing the advantage of its being distributionfree, comparatively insensitive to observational errors and more powerful for skew distributions. These variations, based on the use of expected values of order statistics and quantiles respectively, lead to more extensive computations than Wilcoxon's test, but it will certainly be worthwile to investigate the profits to be gained from this extra work.

Acknowledgement

The computations for this paper have been done by several members of the Statistical Department of the Mathematical Centre, Amsterdam. Thanks are due to A. R. Bloemena and Constance van Eeden, who supervised this work.

References

NOETHER, G. E., The efficiency of some distribution-free tests, Statistica Neerlandica **12** (1958) 63–73

QUENOUILLE, M. H., Tables of random observations from standard distributions, Biometrika **46** (1959) 178–204

Discussion

H. J. PRINS: I would like to make two remarks on Dr. Hemelrijk's paper. The first is that in his conclusions Dr. Hemelrijk stated that for skew distributions (in particular the exponential distribution) the difference between the t-test and Wilcoxon's test was not marked, whereas the t-test was markedly more sensitive to the presence of slipped variates. In general this seems contradictory, as the presence of slipped observations may be regarded as a kind of skewness of the underlying distribution. If the model of slippage used by Dr. Hemelrijk is generalized a little bit, there is in fact *no* difference between the model where *all* observations stem from a skew distribution and the model of a normal distribution with slipped observations. Even the exponential distribution mentioned may be considered thus. The generalized model of slippage mentioned above is the following:

1. n observations are drawn from for example a normal population.
2. For every observation by independant drawings with a certain probability p it is decided whether the observation will be an outlier.
3. If an observation is designated as an outlier, another one is drawn from a specified distribution to indicate the amount of slippage of this observation.

This generalized model seems to describe the occurrence of slippage very well. If this is the case, it is not possible to distinguish between a certain skew distribution and this model.

The second comment is a question. Would not it be useful not to compare the t-test and Wilcoxon's test in case of the exponential distribution, but to compare the two sample test for exponentially distributed observations with Wilcoxon's test in the exponential case?

J. HEMELRIJK: The remark of Mr. PRINS on slippage and skewness is certainly to the point. As it was received after the symposium we had some time to think it over.

The conclusions of the paper were based on the results of the sampling experiment without direct reference to theoretical considerations. But the experiment itself was based on a general feeling that Student's test would probably be more vulnerable with regard to slippage than Wilcoxon's. The experiment confirms this for the special kind of slippage applied. This kind of slippage, where only one error is made — and that one in the sample with the largest median — cannot properly be described in the way indicated by Mr. Prins. Nevertheless, in order to investigate his remark a little further, Ir. A. R. Bloemena was so kind as to compute Pitman's asymptotic relative

efficiency for the case of the exponential distributions used in the experiment. The result was somewhat surprising: he found that this efficiency, for Wilcoxon's test with regard to Student's, was only 0.75, far lower than in the normal case. This is in accordance with Mr. Prins' remark: if Student's test is more sensitive for slippage then it should also be more sensitive for progressive skewness. But the sampling experiment does not confirm this at all. It indicates that, for $m = n = 10$, the difference between both tests in the exponential case is markedly smaller than in the normal and slippage case. This may, of course, be due to m and n being small and it therefore forms an indication that small sample properties may be markedly different from asymptotic ones. On the other hand Noether has proved that for skew distributions with displacement the efficiency of Wilcoxon's test can be much larger than of Student's. In short: Mr. Prins' question underlines the necessity of further research.

As to his second comment about comparing Wilcoxon's test in the exponential case with the optimum test for this case instead of with Student's, this would certainly be interesting but it fell (as yet) outside the scope of the paper.