"Distribution-Free" or "Non-parametric" Methods

Preface [November, 2001, by jh]

The second Edition of Moore and McCabe, which I used until 1997, did not have a section on these methods. The 3rd and 4th edition devote Chapter 14 to these. The chapter is not in the text per se, but rather in a document in pdf format in the CD that comes with the book. In case you don't have the book, I have put their chapter 14 on the web page.

We have also moved on considerably in computing in these last ten years, to the point where many of the non-parametric tests are built into SAS and INSIGHT, and easily programmable in Excel. I give examples in the Resources for Chapter 10.

Despite this ease of calculation, it doesn't hurt to know how to do these tests the old fashioned way -- by pencil and paper,and by table lookup. Doing so also helps you understand what the tests do. In fact, it seems ironic that these tests that were developed partly to avoid tedious calculation, are now them selves computerized!

The material below—a mix of material from Bradford Hill, Peter Armitage and myself—has a few items that M&M Ch14 do not. In particular, it provides the tables for the critical values of the rank statistic -- tables based not on the Gaussian approximation described in M&M14, but on the exact distributions of the rank statistics. At first, it may seem strange to be using Gaussian (parametric!) approximations when the very data being analyzed are decidedly non-Gaussian. But if you think of the ranks as transformations of the data, just like logs are, then it does not seem so strange. And you will see in the material provided under resources that the rank sum distribution in particular has a decidely Gaussian shape *even with very low sample sizes*.

Preamble [1996, by jh, when I prepared the notes below]

Moore and McCabe do not deal with these methods, and so I have assembled some material here to fill the gap. I have taken it from two main sources, the classic Short Textbook of Medical Statistics by the late Sir Bradford Hill, and (more extensively) from the 3rd edition of Statistical Methods for Medical Research by Armitage and Berry [I will refer to them as "A&B"]. I also add some graphical displays and tables of my own to help show some of the detail behind the scenes. Finally, I give some examples from the medical literature by way of exercises.

The material begins with an overview from A&B, then 'cuts to the chase' with one page from Bradford Hill on each of the two most important tests, the Wilcoxon Signed Ranks Test for "1-sample" data and the Wilcoxon Rank Sum Test for 2 independent samples. Then, for those who want to know more, we go back to the material in A&B.

Introduction [taken from §13.1 of A&B]

Some of the statistical methods described earlier in connection with categorical data have involved rather simple assumptions: for example, chi-square methods often test simple hypotheses about the probabilities for various categories—that they are equal, or that they are proportional to certain marginal probabilities. The methods used for quantitative data, in contrast, have relied on relatively complex assumptions about distributional forms—that the random variation is normal, Poisson, etc. These assumptions are often likely to be clearly untrue; to overcome this problem we sometimes argue that methods are *robust*—that is, not very sensitive to non-normality. At other times we may use transformations to make the assumptions more plausible.

Clearly, there would be something to be said for methods which avoided unnecessary distributional assumptions. Such methods, called *distribution-free methods,* exist and are widely used by some statisticians. Standard statistical methods frequently use statistics which in a fairly obvious way estimate certain population parameters; the sample estimate of variance s^2, for example, estimates the population parameter σ^2. In distribution-free methods there is little emphasis on population parameters, since the whole object is to avoid a particular functional form for a population distribution. The hypotheses to be tested usually relate to the nature of the distribution as a whole rather than to the values assumed by some of its parameters. For this reason they are often called *non parametric hypotheses* and the appropriate techniques are often called *non parametric tests* or *methods.* *[often the name given to them in statistical packages ... jh]*

The justification for the use of distribution-free methods will usually be along one of the following]lines.

1 There may be obvious non-normality.

2 There may be possible non-normality, perhaps to a very marked extent, but the sample sizes may be too small to establish whether or not this is so.

[**Note**: its a "Catch-22": when n is large enough to reliably look for non- Gaussian-ness, the CLT means that you don't have to worry about non-Gaussian-ness! when n isn't large enough to check, you cannot call on the CLT... so, with small n, you have to use judgment and imagine what the distribution would look like ..jh]

3 One may seek a rapid statistical technique, perhaps involving little or simple calculation. Many distribution-free methods have this property: J. W. Tukey's epithet 'quick and dirty methods' is often used to describe them.

[I prefer 'quick but not that dirty'... jh]

4 A measurement to be analysed may consist of a number of ordered categories, such as —, –, 0, + and ++ for degrees of clinical improvement; or a number of observations may form a rank order—for example. Patients may be asked to classify six pharmaceutical formulations in order of palatability. In such cases, the investigator may be unwilling to allot a numerical scale, but would wish to use methods which took account of the rank order of the observations. Many distribution-free methods are of this type. The first type of data referred to here, namely ordered categorical data, has already been discussed at some length in Chapter 12. There is, in fact, a close relation between some of the methods described there and those to be discussed in the present chapter.

The methods described in the following sections are merely a few of the most useful distribution-free techniques. These methods have been developed primarily as significance tests and are not always easily adapted for purposes of estimation Nevertheless, the statistics used in the tests can often be said to estimate something, even though the parameter estimated may be of limited interest. Some estimation procedures are therefore described briefly, although the emphasis will be on significance tests. Some of the general issues about the use of distribution free methods are discussed in §13.6.

Fuller accounts of distribution-free methods are given by Siegel and Castellan (1988), who concentrate on significance tests, Lehmann (1975) and Conover (1980)

"FAST TRACK": The Wilcoxon Signed Rank Test Taken from Bradford Hill's Short Textbook of Medical Statistics

In an earlier table , we had the 9 observations of blood pressure **before** and **after** treatment. The **changes** that occurred were shown in column 3. *Ignoring at first the sign of the change,* we can list them in order from smallest to largest, and we can assign to each its appropriate rank number. Thus we have:-

Change	0	2	4	10	11	12	17	29	30
Rank	–	1	2	3	4	5	6	7	8
Sign of Rank		+	+	–	–	–	–	–	–

In the last row we insert the original sign of the change, i.e. whether the blood pressure rose or fell (there was one patient who exhibited no change; this observation is omitted from the test as it provides no sign. Finally we sum the positive ranks (patient's blood pressure rose = (1 + 2) = 3, and the negative ranks (patient's blood pressure fell= (3 + 4 + 5 + 6 + 7 + 8)= 33. We have now to consider these two figures 3 and 33. With 8 observations the total of all the ranks is (1 + 2 + 3 . . . + 8) = 36, and this total can be divided into two sections in 37 different ways (from 0 and 36 to 36 and 0). If the treatment has no effect (the null hypothesis) then we might expect the upward changes in blood pressure that occur by chance to equal the downward changes. In other words, the sums of signed ranks would, in the long run, be –18 and + 18. But with a small number of observations we shall see, by the play of chance, departures from that equality. Is it, then, probable that a difference between the positive and negative sums of 3 to 33 could have arisen by chance? Or is the domination of the negative ranks over the positive 'statistically significant' ? The answer is provided by the table of values given on p. 315.

(jh.. reproduced here .... ------>)

Looking at the values given for 8 paired observations we see that 3,33 is significant at the 0 05 level, i.e. it would occur by chance not more than once in twenty times. Any less extreme division of the 36 ranks, e.g. 4 and 32, would not reach this level of significance and, accepting such a level as necessary, would not contradict the null hypothesis that the treatment had no effect.

If, to take another example, we had had 19 pairs of observations, then the table shows that the sum of all ranks would be 190 and that the division of this total into 46 and 144 would be statistically significant at the 0.05 level and 32 and 158 at the 0.01 level. Any figures between 46 and 144, e.g. 58 and 132, would not be significant.

Table for Wilcoxon Signed Rank Test (Paired Observations)

The sums of signed ranks required for significance at 0 05 and 0 01 levels

Number of **paired** observations showing differences	P = 0.05 (2-sided)	P = 0.01 (2-sided)
6	0, 21	-- --
7	2, 26	-- --
8	3, 33	0, 36
9	5, 40	1, 44
10	8, 47	3, 52
11	10, 56	5, 61
12	13, 65	7, 71
13	17, 74	9, 82
14	21, 84	12, 93
15	25, 95	15, 105
16	29, 107	19, 117
17	34, 119	23, 130
18	40, 131	27, 144
19	46, 144	32, 158
20	52, 158	37, 173
21	58, 173	42, 189
22	65, 188†	48, 205
23	73, 203	54, 222
24	81, 219	61, 239
25	89, 236	68, 257
26	98, 253	75, 276
27	107, 271	83, 295
28	116, 290	91, 315
29	126, 309	100, 335
30	137, 328	109, 356
31	147, 349	118, 378
32	159, 369	128, 400
33	170, 391	138, 423
34	182, 413	148, 447
35	195, 435	159, 471

† For 22 pairs, the required sums for a P value of 0.05 are 65,188. But for sums of 66,187 the P value is 0.05007 so that these sums might also be accepted as significant.

Note from jh: this explains the discrepancy with Table A7 in A&B.

"FAST TRACK": The Wilcoxon Rank Sum Test for Two Groups of Independent Samples of Observations

from Bradford Hill's Short Textbook of Medical Statistics

The above test was applicable to *paired* observations. In comparing t*wo separate sets of observations* the Wilcoxon Rank Sum test is appropriate. In an earlier example (2 **independent** samples) we compared the blood pressures of **9 patients** with a specified illness and of **11** comparable **normal** persons. For the rank sum test, we combine these 20 observations and list them from lowest to highest and give to each its appropriate rank. Thus, with the blood pressures of the 9 patients underlined and those of the normal persons without underlining, we have:-

Values	Ranks	Values	Ranks
98	1	128	11
105	2	129	12
107	3	132	13
110	4	134	14
114	5	136	15
115	6	139	16
123	7.5	140	17
123	7.5	145	18
125	9	154	19
126	10	160	20

In allotting the ranks we have to take into account the fact that there are two identical observations. Maybe, however, these values were rounded off in the process of tabulation and, if we refer back to the original observations, we could find that they were slightly different, e.g. 122.7 and 123.3†. If this be so, then we can give them their appropriate ranks 7 and 8. But if reference back is not possible or still does not distinguish between the two (or more) values, then we must give the average of the ranks involved. [† hard to believe 1 decimal place for BP!]

Thus the two values of 123 occupy the position of ranks 7 and 8 and are each given the average value of 7.5.* We now sum the ranks for the 11 normal persons and for the 9 patients and reach values of 93.5 and 116.5 (there is, of course, no question of a sign). The question at issue is whether such totals as these could have easily arisen by chance. The answer is provided in the table on p. 316-17 *(of Hill's text, reproduced next page)* in which provision is already made for the fact that the two groups were of different sizes and in which the *figures relate to the smaller of the groups.†*

Thus, looking at the row for 9 and 11 observations, we see that in *the smaller of the two groups* we could expect to see a value as low as 68 or as high as 121 only once in 20 trials, i.e. at the 0.05 level. As our observed value is 116.5, i.e. less than 121, the difference between the groups is not quite statistically significant at a level of probability of 0.05.

Two points of importance should be noted. The t test applied to these figures gave a 'statistically significant' result - it was almost precisely at the 0.05 level. The Wilcoxon test has given a 'not statistically significant' result - the total of 116.5 was just below the 0.05 level of 121. Thus it should be realised that two different tests of significance may give different answers - particularly in borderline cases; and secondly, that the t test is in general a more sensitive test for genuine differences than is the rank test.

• This adjustment slightly weakens the validity of the test of significance given in the table on pp. 316-17 and would make it unreliable if there were many identical values.

†With 20 ranks the total is 210. The group of 9 patients is, however 'penalized' in that it lacks two ranks (10 and 11) available to the group of 11 normals. The total to be distributed is, therefore, 210–(10+ 11)=189.

With Gaussian approximation, no table to keep compact, so does not matter which is 'target' group in the analysis.

Table for The Wilcoxon Rank Sum Test for Two Groups of Independent Samples of Observations

from Bradford Hill's Short Textbook of Medical Statistics

Table: The sum of ranks (small or large) required in the smaller of the two groups to establish significance at the 0.05 and 0.01 levels.

n_1, n_2	P = 0.05	P = 0.01	n_1, n_2	P = 0 05	P = 0 01	n_1, n_2	P = 0.05	P = 0.01	n_1, n_2	P = 0 05	P = 0 01
4,4	10,26	- -	*5,19*	34,91	27,98	*7,17*	56,119	47,128	*10,12*	84,146	76 154
4,5	11,29	- -	*5,20*	35,95	28,102	*7,18*	58,124	49,133	*10,13*	88,152	79 161
4,6	12,32	10,34	*5,21*	37,98	29,106	*7,19*	60,129	50,139	*10,14*	91,159	81,169
4,7	13,35	10,38	*5,22*	38,102	29,111	*7,20*	62,134	52,144	*10,15*	94,166	84,176
4,8	14,38	11,41	*5,23*	39,106	30,115	*7,21*	64,139	53,150	*10,16*	97,173	86 184
4,9	14,42	11,45	*5,24*	40,110	31,119	*7,22*	66,144	55,155	*10,17*	100,180	89,191
4,10	15,45	12,48	*5,25*	42,113	32,123	*7,23*	68,149	57,160	*10,18*	103,187	92,198
4,11	l6,48	12,52							*10,19*	107,193	94,206
4,12	17,51	13,55	*6,6*	26,52	23,55	*8,8*	49,87	43,93	*10,20*	110,200	97,213
4,13	18,54	13,59	*6,7*	27,57	24,60	*8,9*	51,93	45,99			
4,14	19,57	14,62	*6,8*	29,61	25,65	*8,10*	53,99	47,105	*11.11*	96,157	87,166
4,15	20,60	15,65	*6,9*	31,65	26,70	*8,11*	55,105	49,111	*11,12*	99,165	90,174
4,16	21,63	15,69	*6,10*	32,70	27,75	*8,12*	58,110	51,117	*1l.13*	103,172	93,182
4,17	21,67	16,72	*6,11*	34,74	28,80	*8,13*	60,116	53,123	*11,14*	106,180	96,190
4,18	22,70	16,76	*6,12*	35,79	30,84	*8,14*	62,122	54,130	*1l.15*	110,187	99,198
4,19	23,73	17,79	*6,13*	37,83	31,89	*8,15*	65,127	56,136	*11,16*	113,195	102,206
4,20	24,76	18,82	*6,14*	38,88	32,94	*8,16*	67,133	58,142	*11,17*	117,202	105,214
4,21	25,79	18,86	*6,15*	40,92	33,99	*8,17*	70,138	60,148	*11,18*	121,209	108,222
4,22	26,82	19,89	*6,16*	42,96	34,104	*8,18*	72,144	62,154	*11,19*	124,217	111,230
4,23	27,85	19,93	*6,17*	43,101	36,108	*8,19*	74 150	64,160			
4,24	27,89	20,96	*6,18*	45,105	37,113	*8,20*	77 155	66,166	*12,12*	115,185	105,195
4,25	28,92	20,100	*6,19*	46,110	38,118	*8,21*	79,161	68,172	*12,13*	119,193	109,203
4,26	29,95	21,103	*6,20*	48,114	39,123	*8,22*	81,167	70,178	*12,14*	123,201	112,212
			6,21	50,118	40,128				*12,15*	127,209	115,221
5,5	17,38	15,40	*6,22*	51,123	42,132	*9,9*	62,109	56,115	*12,16*	131 217	119 229
5,6	18,42	16,44	*6,23*	53,127	43,137	*9,10*	65,115	58,122	*12,17*	135,225	122,238
5,7	20,45	16,49	*6,24*	54,132	44,142	*9,11*	68,121	61,128	*12,18*	139,233	125,247
5,8	21,49	17,53				*9,12*	71,127	63,135			
5,9	22,53	18,57	*7,7*	36,69	32,73	*9,13*	73,134	65,142	*13,13*	136,215	125,226
5,10	23,57	19,61	*7,8*	38,74	34,78	*9,14*	76,140	67,149	*13,14*	141,223	129,235
5,11	24,61	20,65	*7,9*	40,79	35,84	*9,15*	79,146	69,156	*13,15*	145,232	133,244
5,12	26,61	21,69	*7,10*	42,84	37,89	*9,16*	82,152	72,162	*13,16*	150,240	136,254
5,13	27,68	22,73	*7,11*	44,89	38,95	*9,17*	84,159	74,169	*13,17*	154,249	140,263
5,14	28,72	22,78	*7.12*	46,94	40,100	*9,18*	87 165	76,176			
5,1S	29,76	23,82	*7,13*	48,99	41,106	*9,19*	90,171	78,183	*14,14*	160,246	147,259
5,16	30,80	24,86	*7,14*	50,104	43,111	*9,20*	93,177	81,189	*14,15*	164,256	151,269
5,17	32,83	25,90	*7,15*	52,109	44,117	*9,21*	95,184	83,196	*14,16*	169,265	155,279
5,18	33,87	26,94	*7,16*	54,114	46,122						
						10,10	78,132	71,139	*15,15*	184,281	171,294
						10,11	81,139	73,147			

MORE DETAIL: One-sample Distribution-Free tests for location [again from A&B, §13.2 ; parts in sans serif font are mine... jh]

In this section we consider tests of the **null hypothesis that the distribution of a random variable y is symmetric about zero**. If, in some problem, the natural hypothesis to test is that of symmetry about some other value, μ, all that need be done is to subtract μ from each observation; the test for symmetry about zero can then be used. The need to test for symmetry about zero commonly arises with paired comparisons of two treatments, when the variable y is the difference between two paired readings. {note from jh: I have changed his *x* to my *y* }

The normal-theory test for this hypothesis is, of course, the one-sample t test, and we shall illustrate the present methods by reference to the "increase in sleep" data which were analysed by a paired t test in our M&M §7.1.

The sign test (see M&M p519-521)

Suppose the observations in a sample of size n are y_1, y_2, . . ., y_n, and that of these r are positive and s negative. Some values of x may be exactly zero, and these would not be counted with either the positives or the negatives. The sum r + s may therefore be less than n, and will be denoted by n'. On the null hypothesis positive and negative values of y are equally likely.

Both r and s therefore follow a binomial distribution with parameters n' and π=1/2. Excessively high or low values of r (or equivalently, of s) can be tested exactly from tables of the binomial distribution....

See additional notes, and a ready to use Table for the Sign Test, under Resources for Ch 14.

The signed rank sum test

Notice that textbooks are not consistent in their use of terms for the 1-sample test: some call it *the signed rank sum test* Others leave out the 'sum' and -- just like Hill does -- call it *the signed rank test*. **To avoid confusion as to whether one is using this or the 2-sample test for independent samples, you might do better by calling this one the 'the non-parametric analog of the -sample t-test' and the other 'the non-parametric analog of the t-test for 2 independent samples'. (jh)**

The sign test clearly loses something by ignoring all information about the numerical magnitudes of the observations other than their sign. If a high proportion of the numerically large observations were positive this would strengthen the evidence that the distribution was asymmetric about zero, and it seems reason able to try to take this evidence into account. Wilcoxon's (1945) signed rank sum test works as follows. The observations are put in ascending order of magnitude ignoring the sign, and given the ranks 1 to *n'* (zero values being ignored as in the sign test). Let T_+ be the sum of the ranks of the positive values and T_- that of the negative. On the null hypothesis T_+ and T_- would not be expected to differ greatly; their sum $T_+ + T_-$ is n'(n' + 1)/2, so an appropriate test would consist in evaluating the probability of a value of, say, T_+ equal to or more extreme than that observed. Table A7 [in A&B] gives critical

values for the *smaller* of T_+ and T_-, for two-sided tests at the 5% and 1% levels, for n' up to 25. [In the material here I have taken the more extensive table from Hill's text (jh)] The distribution is tabulated fully for *n' up to* 20 by Lehmann (1975, Table H), and other percentiles are given in the Geigy Scientific Tables (19~2, p. 163). For larger values of n', T_+ and T_- are approximately normally distributed with variance n'(n' + 1)(2n' + 1)/24, and a standardized normal deviate, with continuity correction, is given by

$$\frac{| T+ - 0.25n'(n' + 1) | - 0.5}{\sqrt{[n'\{n' + 1\}\{2n' + 1\}/24]}} \qquad (13.1)$$

If some of the observations are numerically equal, they are given tied ranks equal to the mean of the ranks which they otherwise would have received. This feature reduces the variance of T_+ by $(t^3—t)/48$ for each group of t tied ranks and the critical values shown in Table A7 are somewhat conservative (i.e. the result is somewhat more significant than the table suggests).

Example(jh): We go back to our example in §7.1 on the 10 differences in sleep when taking an active versus a placebo medication. Remember that there were r=8 positives' among n'=10 non-zero differences. The differences were

```
0.9  -0.9  4.3  2.9  1.2  3.0  2.7  0.6  3.6  -0.5
```

Ignoring at first the sign of the change, we can give them each a rank according to their absolute magnitudes

```
3.5   3.5   10   7    5    8    6    2    9     1
```

We use the signs of the original observations

```
 +     –     +   +    +    +    +    +    +     –
```

and then add the ranks associated with positive and negative signs to get

T+ = 3.5 + 10 + 7 + 5 + 8 + 6 + 2 + 9 = **50.5**

and

T– = 3.5 + 1 = **4.5**

[aside: check that T+ and T– add to 55, the sum of the first 10 natural numbers, or 10 times the average rank of 5.5; or you can use this fact to do the least work, in this case by calculating T– first]

From Table A7, for n' = 10, the 5% point for the split of the 55 is (8,47) or (47,8) i.e. under H_0, an (8,47) split or anything more extreme [(48,7),(49,6),.. (55,0) on the one side or (7,48), (6,49), ... (0,55) on the other] should happen less than 1 time in 20. The observed split is greater than this critical value, and so the 2-tail p-value is less than 0.05.

How much less? From Table A7, for n' = 10, the 1% point

for the split of the 55 is (3,52) or (52,3) so our observed value of (50.5,4.5) is between the 0.05 and 0.01 critical values, closer to the 0.01 .

How about calculating exactly the probability of a split as extreme or more extreme, on one side or the other side, than that observed.

For small n, it is possible to calculate out the probabilities for a T+ of 0, 1, 2, ... , 54, 55. These are given as the last diagonal in the spreadsheet grid I have entitled "Excerpts from the distribution of the signed rank statistic" (in a separate file under Resources). If you can consult the notes that accompany it, you will be able to determine that the following chances (out of a total of 210 = 1024] of getting various T+ values:

```
T+     46  47  48  49  50  51  52  53  54  55
Prob:   8   6   5   4   3   2   2   1   1   1 / 1024
```

so that the chances are:

```
 1/1024 of a T+ of 55
 2/1024 of a T+ of 54 or more
 3/1024 of a T+ of 53 or more
 5/1024 of a T+ of 52 or more
 7/1024 of a T+ of 51 or more
10/1024 of a T+ of 50 or more
14/1024 of a T+ of 49 or more
19/1024 of a T+ of 48 or more
25/1024 of a T+ of 47 or more
34/1024 of a T+ of 46 or more
```

[Digression: We can now check the accuracy of Hill's tables, where the critical value for 5% 2-sided is 47. The probability of 46 or more on the one side or 9 or fewer on the other is 2 x 34/1024 = 0.0664. The probability of 47 or more on the one side or 8 or fewer on the other is 2 x 25/1024 = 0.0488. For the 0.01 level, he gives the critical split as (3,52) or (52,3). Again, the probability of 51 or more on the one side or 4 or fewer on the other is 2x7/1024 = 0.0137. The probability of 52 or more on the one side or 3 or fewer on the other is 2x5/1024 = 0.0097, So in these 2 checks at least, the tables are as advertised]

Coming back to the p-value for 50.5. Obviously, there is no entry corresponding exactly to 50.5 but if we interpolate between the tail of 10/1024 for 50 and 7/1024 for 51, we can say that the 2-tail p-value is 2x8.5/1024 or 0.0166.

For the large-sample approximation to the test, we calculate

$$E(T+) = 0.25(10)(11) = 27.5,$$

$$\text{var}(T+) = 10(11)(21)/24 - (1/48)[2^3 - 2]$$

$$= 96.25 - 0.125 = 96.125$$

so that the standardized normal deviate is

$$z = \frac{|\,50.5 - 27.5\,| - 0.5}{\sqrt{96.125}} = 2.29$$

so that

1-tail p is P[Z>2.29] = 0.0110,

and 2-tail p is P[Z>2.29] = 2 x 0.0110 = 0.0220.

The signed rank sum test cont'd

Interval Estimation :
using distribution-free tests to construct CI's

Suppose the observations (or differences, in the case of a paired comparison as in Example 13.2) are distributed symmetrically not about zero, as specified by the null hypothesis, but about some other value, μ. How can we best estimate μ? One obvious suggestion is the sample mean. Another is the sample median, which, if subtracted from each observation, would give the null expectation in the sign test, since there would be equal numbers of positive and negative differences. A somewhat better suggestion is related to the signed rank test. We could choose that value $\hat{\mu}$, which if subtracted from each observation, would give the null expectation in the signed rank test. It is not difficult to see that the test statistic T+ is the number of positive values among the pair means, which are formed by taking the mean of each pair of observations (including each observation with itself). The estimate $\hat{\mu}$ is then the median of these pair means.

Confidence limits for μ are the values which, if subtracted from each observation, just give a significantly high or low test result. For this purpose all n readings may be used. **The limits may be obtained by ranking the n(n + 1) pair means, and taking the values whose ranks are one greater than the appropriate entry in Table A7 (p. 577), and the symmetric rank obtained by subtracting this from n(n + 1)/2 + 1. That is, one excludes the tabulated number of observations from each end of the ranked series.**

The procedure is illustrated below. Because of the discreteness of the ranking, the confidence coefficient is somewhat greater than the nominal value (e.g. greater than 95% for the limits obtained from the entries for 0 05 in Table A7 [or the table from Hill]). If there are substantial ties in the data, a further widening of the confidence coefficient takes place.

Example The 10 differences from the sleep data used earlier in this material for the signed rank sum test are shown in the following table, arranged in ascending order in both rows and columns. They give the following 55 [=0.5 x 10 x 11] pair means:

	−0.9	−0.5	+0.6	+0.9	+1.2	+2.7	+2.9	+3.0	+3.6	+4.3
−0.9	−0.9	−0.7	−0.15	0	+0.15	+0.9	+1.0	+1.05	+1.35	+1.7
−0.5		−0.5	+0.05	+0.2	+0.35	+1.1	+1.2	+1.25	+1.55	+1.9
0.6			+0.6	+0.75	+0.9	+1.65	+1.75	+1.8	+2.1	+2.45
0.9				+0.9	+1.15	+1.8	+1.9	+1.95	+2.25	+2.6
1.2					+1.2	+1.95	+2.05	+2.1	+2.4	+2.75
2.7						+2.7	+2.8	+2.85	+3.15	+3.5
2.9							+2.9	+2.95	+3.25	+3.6
3.0								+3.0	+3.3	+3.65
3.6									+3.6	+3.95
4.3										+4.3

Note that the numbers of positive and negative pair means (counting zero values as contributing l/2 to each sum) are 50.5 and 4.5, respectively, agreeing with the values of T+ and T– obtained earlier.

The estimate $\hat{\mu}$ is the median value of the 55 pair means, namely 1.8. [I'd welcome advice on how to get to this quickly].

For a 95% confidence interval for μ[increase in sleep], note that the entry in Table 7A for n=10 and P=0.05 is 8 so one looks for the 9th pair mean and the 56-9 = 47th pair mean as the lower and upper limits. In this example, they are +0.35 and 3.15 respectively. For comparison, the t distribution use for these data in §7.1 gave limits of +0.5 to +3.0 hours.

Application: study of Chalasia Chair for GastroEsophageal Reflux --- analysis (cf NEJM 1983 309:760-763 see §7)

data

infant	GER time chair	prone	diff	*rank of |diff|*
1	14	17	-3	*2*
2	62	16	46	*9*
3	18	12	6	*4*
4	9	3	6	*4*
5	49	16	33	*8*
6	18	4	14	*6*
7	10	8	2	*1*
8	30	1	29	*7*
9	44	38	6	*4*

descriptive statistics

	CHAIR	PRONE	DIFF
N	9	9	9
MIN	9	1	-3
MAX	62	38.0	46
MEAN	28.2	12.8	15.4
SDEV	19.2	11.2	16.6
SE	6.4	3.7	5.5

PAIRED SAMPLES T-TEST
CHAIR VS PRONE WITH 9 CASES

MEAN DIFFERENCE = 15.444
SD DIFFERENCE = 16.644
T=2.784 DF=8 PROB = .024

SIGN TEST RESULTS

COUNTS OF DIFFERENCES
(ROW VARIABLE > COLUMN)

	CHAIR	PRONE
CHAIR	0	8
PRONE	1	0

TWO-SIDED PROBABILITIES
FOR EACH PAIR OF VARIABLES

	CHAIR	PRONE
CHAIR	1.000	
PRONE	.039	1.000

WILCOXON SIGNED RANKS TEST

COUNTS OF DIFFERENCES
(ROW VARIABLE > COLUMN)

	CHAIR	PRONE
CHAIR	0	8
PRONE	1	0

TWO-SIDED PROBABILITIES
USING NORMAL APPROXIMATION

	CHAIR	PRONE
CHAIR	1.000	
PRONE	.015	1.000

From Hill Table for Signed Ranks

$\sum-$ = 2; $\sum+$ = 43. 9 diffs

If (1,44) split ==> P_{2sided} <0.01

If (5,40) split ==> P_{2sided} <0.05

Our observed split (2,43) is between the 0.05 and 0.01 cutoffs.

Calculating P-value more precisely ... From "Excerpts from Distribution of Signed Ranks statistic" Spreadsheet (undr Resources)

n =9 ;

$P(\sum- = 0, \sum+ = 45) = 1 / 512$*
$P(\sum- = 1, \sum+ = 44) = 1 / 512$
$P(\sum- = 2, \sum+ = 43) = 1 / 512$

P(observed or more extreme) 3 / 512

P(observed or more extreme other tail
ie {45,0},{44,1} and {43,2}) 3 / 512

Total ie $P_{2tailed}$ 6/ 512

= **0.0117**

* 46 frequencies 1,..,23,..1 on diagonal n=9 add to 2^9 = 512

MORE DETAIL: Two-sample Distribution-Free tests for location [again from A&B, §13.2 ; parts in sans serif font are mine..jh]

Suppose we have two groups of observations: a random sample of n_1 observations, x_i, from population X and a random sample of n_2 observations, y_j, from population Y. The null hypothesis to be tested is that the distribution of x in population X is exactly the same as that of y in population Y. We should like the test to be sensitive to situations in which the two distributions differ primarily in location, so that x tends to be greater (or less) than y. The normal-theory test is the two-sample (unpaired) t test described in §4 of A&B [§7 of M&M]. Three distribution-free tests in common usage are all essentially equivalent to each other. They are described briefly here.

The Mann-Whitney U test

The observations are ranked together in order of increasing magnitude. There are n_1n_2 pairs (x_i, y_j); of these,

U_{XY} is the number of pairs for which $x_i < y_j$,

and

U_{YX} is the number of pairs for which $x_i > y_j$.

Any pairs for which $x_i > y_j$ count 1/2 a unit towards both U_{XY} and U_{YX} Either of these statistics may be used for a test, with exactly equivalent results. Using U_{YX}, for instance, the statistic must lie between 0 and n_1n_2. On the null hypothesis its expectation is $n_1n_2/2$. High values will suggest a difference between the distributions, with x tending to take higher values than y. Conversely, low values of U_{YX} suggest that x tends to be < y.

Wilcoxon's rank sum test

Again there are two equivalent statistics:

T_1 is the sum of the ranks of the x_i's;

T_2 is the sum of the ranks of the y_j's.

Low values assume low ranks (i.e. rank 1 is allotted to the smallest value). Any group of tied ranks is allotted the midrank of the group.

The smallest value which T_1 can take arises when all the x's are less than all the y's; then $T_1 = n_1(n_1 + 1)/2$. The maximum value possible for T_1 arises when all x's are greater than all y's; then $T_1 = n_1n_2 + n_1(n_1 + 1)/2$. The null expectation of T_1 is $n_1(n_1 + n_2 + 1)/2$.

Interrelationships between tests

There are, first, two relationships between the two Mann-Whitney statistics and between the two Wilcoxon statistics:

$$U_{XY} + U_{YX} = n_1n_2 , \qquad (13.3)$$

$$T_1 + T_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2. \qquad (13.4)$$

These show that tests based on either of two statistics in each pair are equivalent; given T_1 and the two sample sizes, for example, T_2 can immediately be calculated from (13.4).

Secondly, the two tests are interrelated by the following formulae:

$$U_{YX} = T_1 - n_1(n_1 + n_2 + 1)/2, \qquad (13.5)$$

$$U_{XY} = T_2 - n_2(n_1 + n_2 + 1)/2, \qquad (13.6)$$

The two tests are exactly equivalent. From (13.5) for instance, the probability of observing a value of T_1 greater than or equal to that observed is exactly equal to the probability of a value of U_{YX} greater than or equal to that observed. Significance tests based on T_1 and U_{YX} will therefore yield exactly the same significance level. The choice between these tests depends purely on familiarity with a particular form of computation and accessibility of tables.

The probability distributions (under the null) of the various statistics are independent of the distributions of x and y. They have been tabulated for small and moderate sample sizes, for situations in which there are no ties. Table A8 gives critical values for T_1 (the samples being labelled so that $n_1 \le n_2$). The table provides for two-sided tests at the 5% and 1% levels, for n_1 and n_2 up to 15. More extensive tables are given in the Geigy Scientific Tables (1982, pp 156-162), and the exact distribution (in terms of U_{XY}) is given by Lehmann (1975, Table B). (Table from Hill goes to $n_1+n_2=30$) Beyond the range of Table A8, the normal approximation based on the variance formulae of Table 13.1 is adequate unless the smaller of n_1 and n_2 is less than 4.

When there are ties, the variance formulae are modified as shown in Table 13.1. The summations in the formulae are taken over all groups of tied observations, t being the number of observations in a particular group. As with the signed rank sum test, the tables of critical values are somewhat conservative in the presence of ties.

Table 13.1[from A&B]**. Some properties of 2 equivalent two-sample distribution-free tests**

	Bounds		**Sampling Distribution**		
				Variance	
			Mean	**No ties**	**Ties**
	xxyy*	yyxx**			
Mann-Whitney U test					
U_{XY}	$n_1\ n_2$	0	$\frac{n_1\ n_2}{2}$	$\frac{n_1\ n_2(n+1)}{12}$	$\frac{n_1\ n_2}{12n(n+1)}[n^3 - n - \sum(t^3 - t)]$
# pairs: $x_i < y_j$					
U_{YX}	0	$n_1\ n_2$	$\frac{n_1\ n_2}{2}$	" "	" "
# pairs: $x_i > y_j$.					
Wilcoxon rank sum test					
T_1	†	††	$\frac{n_1\ (n+1)}{2}$	" "	" "
∑ranks for x					
T_2	¶	¶¶	$\frac{n_2\ (n+1)}{2}$	" "	" "
∑ranks for y					

---

xxyy* : all x < all y yyxx** : all y < all x ; $n = n_1 + n_2$ † $n_1\ (n_1 + 1)/2$ †† $n_1\ n_2 + n_1\ (n_1 + 1)/2$ ¶ $n_1\ n_2 + (n_2 + 1)/2$

¶¶ $n_2\ (n_2 + 1)/2$

[If using Gaussian approximation, can use either larger or smaller sample size, since no special table, and thus no imperative to keep table small]

Application: Investigation of methods to reduce dropouts from exercise classes

- **Study Design:**
  8 exercise classes sat U de M, 4 of them randomly assigned to receive weekly counselling by sports psychologist on how to 'hang in there'

- **Data**

Class	Type	Average no. of sessions attended	*Rank*
1	1(e)	11.1	*6*
2	1(e)	12.2	*8*
3	1(e)	9.4	*2**
4	1(e)	11.7	*7*
5	2(c)	9.6	*3*
6	2(c)	9.2	*1*
7	2(c)	10.3	*5*
8	2(c)	9.7	*4*

- **Descriptive statistics**

"Y" variable = ATTENDED

	e	c
N OF CASES	4	4
MINIMUM	9.400	9.200
MAXIMUM	12.200	10.300
MEAN	11.100	9.700
STANDARD DEV	1.219	0.455
STD. ERROR	0.610	0.227
variance	1.49	0.21
average variance	0.85	

- **t-test**

$$t_6 = \frac{11.1 - 9.7}{\sqrt{0.85\{\frac{1}{4} + \frac{1}{4}\}}}$$

$$= \frac{1.4}{0.65} = 2.15 \quad [t_{6,0.05} = 2.447 \quad \text{so } \partial \text{ is "ns"}]$$

$$t^2 = 2.15^2 = 4.63$$

- **ANALYSIS OF VARIANCE**

```
DEP VAR:ATTENDED
N:                    8
MULTIPLE R:           .660
SQUARED MULTIPLE R:   .436
```

SOURCE	SS	DF	MS	F	P
TYPE	3.92	1	3.920	4.63	0.075
ERROR	5.08	6	0.847		

- **NON-PARAMETRIC ANALYSIS†**

```
KRUSKAL-WALLIS 1-WAY ANOVA
  DEPENDENT VARIABLE IS ATTENDED
  GROUPING VARIABLE IS     TYPE
```

GROUP	COUNT	RANK SUM
1(e)	4	23
2(c)	4	13

MANN-WHITNEY U TEST STATISTIC = 13 [3+1+5+4]
PROBABILITY IS 0.149
CHI-SQ APPROXN¶ = 2.083 WITH 1 DF

---

† same as Wilcoxon & Mann Whitney when 2 groups

¶ $\sqrt{\text{CHI-SQ}} = Z = \frac{(13 - 18)}{\sqrt{4x4x9/12}} = -1.44$

μ & SD from table 13.1

* class met at 8AM: *[See "statistics for random assignment of intact classrooms to treatments" in Campbell DT and Stanley JC "Experimental and Quasi-experimental Designs for Research" Rand McNally Chicago 1963, page 23] Although a total of 8x25 students in the 8 classes, giving 100 for E and 100 for C, the 'real' n's are n_1=4 and n_2=4.*

Application: Analyses of data in "Routine ultrasonography in utero and school performance at age 8-9 years"

Routine ultrasonagraphy in utero and school performance at age 8–9 years

Most fetuses in developed countries are exposed in utero to diagnostic ultrasound examination. Many pregnant women express concern about whether the procedure harms the fetus. Since most routine ultrasound examinations are done at weeks 16–22, when the fetal brain is developing rapidly, effects on neuronal migration are possible. We have sought an association between routine ultrasonography in utero and reading and writing skills among children in primary school.

At the age of 8 or 9 years, children of women who had taken part in two randomised, controlled trials of routine ultrasonography during pregnancy were followed-up. The women had attended the clinics of 60 general practitioners in central Norway during 1979–81. The analysis of outcome was by intention to treat: 92% of the "screened" group had been exposed to ultrasound screening at weeks 16–22, and 95% of controls had not been so exposed, but there was some overlap. 2428 singletons were eligible for follow-up, and the school performance of 2011 children (83%) was assessed by their teachers on a scale of 1-7; the teachers were unaware of the teachers were unaware of ultrasound exposure status. A subgroup of 603 children underwent specific tests for dyslexia. There were no statistically significant differences between children screened with ultrasound and controls in the teacher-reported school performance (scores for reading, spelling, arithmetic, or overall performance). Results from the dyslexia test sample showed no differences between screened children and controls in reading, spelling, and intelligence scores, or in discrepancy scores between intelligence and reading or spelling. The test results classified 21 of the 309 screened children (7% [95% confidence interval 3–10%) and 26 of the 294 controls (9% [4-12%]) as dyslexic.

The risk of having poor skills in reading and writing was no greater for children whose mothers had been offered routine ultrasonography than for those whose mothers had not been offered the procedure.

K A Salvesen et al *Lancet 1992; 339: 85–89.*

TABLE I—COMPARISON OF CHILDREN IN STUDY WITH THOSE NOT AVAILABLE

	Mother lost to follow-up	Did not respond to q'aire	No teacher assessment	Final study group
n	40	261	150	2011
Mean maternal age (yr) at pregnancy	26	26	25	26
% of mothers with education of:				
6-9 yr	38	46	38	39
9-12 yr	47	41	49	48
> 12 yr	15	13	13	13
% non-smoking mothers	76	54	65	63
Mean(SD) no. of ultrasound examinations				
Screened group	1.3(1.6)	2.0(1.0)	2.3(0.8)	2 3(0.9)
Control group	0.1(0.4)	0.2(0.4)	0.5(1.1)	0.3(0.8)
% ma/e	43	55	61	50
% left-handed	. .	. .	13	9
% with family history of dyslexia	. .	. .	23	13
% of children with history of allergies.	. .	. .	14	21

Application: Analyses of data in "Routine ultrasonography in utero and school performance at age 8-9 years" ... continued...

TABLE II—TEACHER ASSESSMENTS OF SCHOOL PERFORMANCE†

No of children (n= 2011)

	Oral reading		Reading comprehension		Arithmetic		Overall performance	
	Scr	Ctl	Scr	Ctl	Scr	Ctl	Scr	Ctl
1	15	14	11	21	9	9	7	8
2	55	53	33	71	35	39	33	33
3	102	104	83	103	81	75	81	85
4	153	186	174	1259	209	205	226	238
5	180	182	170	1195	210	236	217	230
6	236	202	237	2197	261	260	248	229
7	266	245	297	2135	199	165	194	164
tot	1007	986	1005	9981	1004	989	1006	987
?	8	10	10	15	11	7	9	9

† I've omitted "spelling" because of lack of space... [jh]

TABLE III—DYSLEXIA TEST RESULTS FOR 309 SCREENED CHILDREN AND 294 CONTROLS

	Mean (SD) Z score Screened	Control	P
Reading	0 02 (1 04)	- 0 02 (0-96)	0.6
Spelling	0 005 (1 01)	- 0 005 (0 99)	0.9
Intelligence	0.19 (1.01)	- 0-20 (0 99)	0.6
DRI	0 01 (1.05)	- 0 01 (0 95)	0 7
DSI	- 0 004 (1 00)	0 004 (1 00)	0 9

DRI = discrepancy score (reading-intelligence);
DSI = discrepancy score (spelling – intelligence)

Setting up data for analysis: *.. see schema ---->*
Data are in 'grouped format'. Instead of creating file with 1 entry/child, create1 1 record per "type of child" and use the FREQ statement in SAS to tell it the 'multiplicity' of each record. e.g. for overall performance", the datafile had 3 variables (group perf_score number) and 14 'observations'

KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE *(on ranks)*

(KRUSKAL-WALLIS test is the multi-group version of the Wilcoxon Rank Sum Test for 2 independent samples)

overall performance (1993 children)

	Screened	**Control**
n	**1006**	**987**
RANK SUM	**1024712**	**962309**
AVERAGE RANK	**1018.6**	**975.0**

MANN-WHITNEY U	χ^2 APPROXN* ($z^2=\chi^2_{1df}$)	PROB
518191	**2.994**	**0.084**

Check:

$\mu[\sum \text{ranks in ctl group}] = 987\text{x}(987+1006+1)/2 = 984039$
$\text{Var}[\sum \text{ranks in ctl group}] = 987\text{x}1006\text{x}(1994)/12 = 12844.86^2$

$$Z = \frac{\sum \text{ranks in ctl group} - \mu}{\sqrt{\text{var}[\sum \text{ranks in ctl group}]}} = \frac{962309 - 984039}{12844.86} = -1.69$$

so = $\{-1.69\}^2 = 2.86$ -- close to χ^2 -- discrepancy may be use of variance corrected for tied ranks.

Interpretation of U statistic

Compare performance of a screened and a control child:
Total of 1006 x 987 = 992922 possible comparisons

Prob(screened > control) = 518191/992922 = 0.52 = 52%

	Screened	*1*	*7*
	Screened	*2*	*33*
data file	*...*	*..*	*..*
(3 variables;	*Screened*	*7*	*194*
14 obsns)	*Control*	*1*	*8*
	...	*..*	*..*
	Control	*7*	*164*

INDEPENDENT SAMPLES T-TEST ON (1–7) **SCORE** for OVERALL

Screened AVE (SD)	**Control** AVE (SD)	SEPARATE VARIANCES T DF=1990.9	POOLED VARIANCES T DF=1991
5.120 (1.4)	**5.018 (1.4)**	T=1.637 P=0.102	T=1.637 P=0.102

Other measures ----- RANKS ------ -- SCORES-

	χ^2 Approx to MANN-WHITNEY U	PROB	PROB(t-test)
oral reading	2.470	0.116	0.159
reading comprn	0.708	0.400	0.384
spelling	2.921	0.087	0.093
arithmetic	1.209	0.272	0.314

Artificial e.g. larger Δ between groups: spelling scores of screened vs. reading comprehension scores of control

"score" (total = 1980)

	Screened	**Control**
n	1001	979
RANK SUM	903374.5	1057815.5
AVE RANK	902.5	1080.5

MANN- WHITNEY U	χ^2 APPROXN	PROB
401873.5	49.904	0.000

U statistic

Total of 1001 x 979 = 979979 possible comparisons

Prob(screened > control)=401873.5/979979 =0.40 = 40%

INDEPENDENT SAMPLES T-TEST ON (1–7) SCORE

	Screened	**Control**
AVE(SD)	4.81**(**1.54**)**	5.28**(**1.52**)**

SEPARATE-VARIANCES T DF=1977.7 T=-6.854 P=0.000

POOLED-VARIANCES T DF=1978 T=-6.853* P=0.000

TEST HOMOGENEITY VARIANCES: χ^2 = 0.11 PROB = 0.741
POOLED WITHIN GROUPS STANDARD DEVIATION = 1.532**

ANALYSIS OF VARIANCE

N: 1980; MULTIPLE R: 0.152; SQUARED MULTIPLE R: 0.023

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
GROUP	110.191	1	110.191	46.962*	0.000
ERROR	4641.154	1978	2.346**		

*note 46.962 = $(-6.853)^2$ **note 2.346 = $(1.532)^2$

Comparison of SEVERAL groups [again from A&B, §13.4 ; parts in sans serif font are mine..jh]

Related groups: Friedman's test

Suppose we have more than two groups of observations and the data are also classified by a block structure so that the data form a two-way classification of the type considered in §8.1, where the two-way analysis of variance of a randomized block design was described. Suppose that there are t treatments and b blocks. A distribution-free test for such a situation was given by Friedman (1937) and this test is a generalization of the sign test to more than two groups. The test is based on ranking the values within each block. The test procedure can best be explained by considering a two-way analysis of variance of the ranks, using the formulae of §8.1. In such an analysis the sums of squares and their degrees of freedom, may be written as follows:

	SSq	DF
Blocks	0	0
Treatments	S_{tr}	t-1
Residual	S_{res}	(b–1)(t–1)
Total	S_{tot}	b(t–1)

Both the sum of squares and the degrees of freedom for blocks are zero because the sum of the ranks is the same for every block, namely t(t + 1)/2. In the calculation of the sums of squares the correction term (T^2/N in Table 8.2) is $bt(t + 1)^2/4$ The usual form of the Friedman test statistic is

$$T_1 = \frac{b\{t—1\}S_{tr}}{S_{tot}} = \qquad (13.7)$$

which is distributed approximately as χ^2 with (t–1) df.. This statistic is the ratio of the treatment SSq to the Total MSq in the analysis of variance. A somewhat preferable test statistic is analogous to the usual variance ratio in the analysis of variance, i.e. the ratio of the Treatment MSq to the Residual MSq:

$$T_2 = \frac{\{b—1\}S_{tr}}{S_{tot}—S_{tr}} \qquad (13.8)$$

which is distributed approximately as F, with t–1 and (t–1)(b–1) df.
When there are no ties the Total SSq, S_{tot} can be calculated directly as b(t + 1)(2t + 1)/6, and the formulae for the test statistics are often written in terms that make use of this expression. When t = 2 the test statistic T_1 is identical to the sign test statistic without a continuity correction.

Table 13.3 shows the data given in Table 8.3 with the values within each block ranked.
The effect of treatments is highly significant by either test, in agreement with the analysis of Example 8.1. Note that T_2 is more significant than T_l; this will generally be true when the effect is in any case highly significant, because the Residual MSq used as the basis T_2 will then be substantially smaller than the Total MSq used in T_1 . This is the reason for the general preference for the second of the two tests.

Table 13.3. Clotting times of plasma from eight subjects, treated by four methods. after ranking the four values within each subject

	Treatment				
Subject	1	2	3	4	Total
1	1	2	3	4	10
2	1	4	2	3	10
3	3	1	4	3	10
4	2	1	3	4	10
5	2	1	3.5	3.5	10
6	1	3	2	4	10
7	1	2	3	4	10
8	1	2	3	4	10
Total	11.0	16.0	23.5	29.5	80

Correction term = $80^2/32$= 200,

Between-Treatments SSq,
$S_{tr} = \{11.0^2 + 16.0^2 + 23.5^2 + 29.5^2\}/8—200 = 24.937$

Total SSq, $S_{tot} = 1^2 + 2^2 + . + 4^2 –200 = 39.5$,

$T_1 = 8 \times 3 \times 24.9375/39.5 = 15.15$ (as $\chi^2_{(3)}$; P=0.002)

$T_2 = 7x24.9375/14.5625 = 11.99$ (as $F_{3,21}$: P< 0.001)

In some studies treatments may be compared within strata or blocks, but with more than one observation per treatment in each stratum, and the numbers of replicates may be unbalanced, as in Example 8.9. This situation is referred to later in this section.

Independent groups: the Kruskal-Wallis test

Suppose we have more than two groups of observations and the data form a one-way classification of the type considered in §7.1, where the one-way analysis of variance was described. Suppose that there are t groups. A distribution-free test for such a situation would be a generalization of the Mann Whitney or Wilcoxon rank sum test to more than two groups. A generalization was given by Kruskal and Wallis (1952). This test is based on ranking all the values and then the test proceeds by a method which has similarities with a one-way analysis of variance on the ranks.

Suppose that there are n_i observations for group i, and let $N = \sum n_i$. The observations are ranked from 1 to N, and the ranks are subjected to a one-way analysis of variance. Let T_i be the sum of the ranks in group i. Denote the corrected sum of squares for groups by S_{tr}, and the corrected total sum of squares by S_{tot}. In these calculations the correction term is given by $0.25N(N + 1)^2$. The test statistic is then calculated as

$$T = \frac{\{N-1\}S_{tr}}{S_{tot}} \qquad (13.9)$$

distributed approximately as $\chi^2(t-1)$. If there are no ties, Stot can be evaluated directly and the formula for the test statistic simplifies to

$$T = \frac{12}{N\{N+1\}} \sum T_i^2 / n_i - 3(N+1) \qquad (13.10)$$

Example: The Prime Minister, Mrs Thatcher, & dementia

[Fuller GN and Meeran K. Lancet Jume 1, 1991; 337(8753) p 1362.

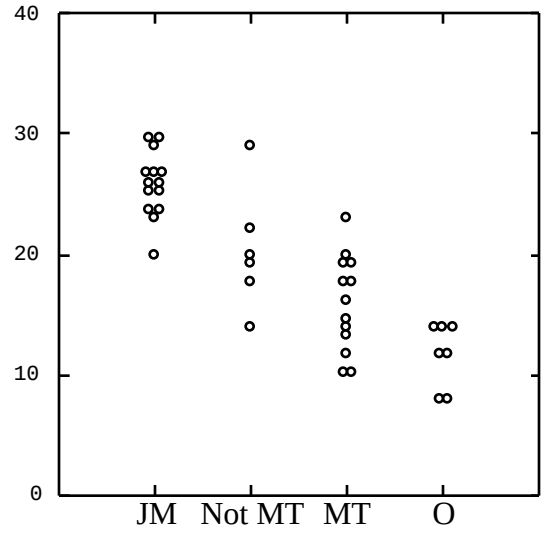
Asking the name of the Prime Minister has frequently been used as part of bedside testing of higher function. However, after five years in office Mrs Thatcher was found to be significantly more memorable than her predecessors,[1] rendering the test less useful. After her resignation, after eleven years, Le Fanu suggested that the test could be reintroduced.[2] We studied 40 patients over 70 years (mean age 80, range 70-98) admitted acutely to hospital during a one-week period a month after Mrs Thatcher resigned. They were asked the name of the Prime Minister, and their answers were classed as Mr John Major (JM), not Mrs Thatcher (Not MT), Mrs Thatcher (MT), or other (O). The mini mental state (MMS) test,[3] a simple reproducible screen of higher mental function, was also administered in a standardised way by a single observer. Results were compared by Mann-Whitney U test. Patients who knew the name of the new Prime Minister had significantly higher MMS scores than any of the other three groups (JM > Not MT, p < 0 02; JM > MT, p < 0 001). Those who knew it was not Mrs Thatcher had higher scores, though not significantly, than those who thought it still was, and both groups fared better than those who had suggested someone else such as Churchill or Macmillan (MT and Not MT < O, p < 0.04; JM > O p < 0 001).
Only 1 patient with an MMS score below 23 knew John Major was the Prime Minister and only 1 patient with a score above this did not. Previous studies found that MMS scores of less than 23 are associated with dementia.[3] The name of the Prime Minister, having been redundant during much of Mrs Thatcher's tern of office, seems to have re-emerged as a simple, powerful tool in screening for dementia. G. N. FULLER, K. MEERAN, Department of Neurology. Charing Cross Hospital, London

1. Deary IJ, Wessley S, Farrell M. Dementia and Mrs Thatcher. Br Med J 1985; 291: 1786.
2. Le Fanu J. Medicine man. Sunday Telegraph Review Dec 23, 1990: vii.
3. Dick JPR, Guiloff RJ, Stewart A, at al. Mini mental state examination in neurological patients. J Neurol Neurosurg Psychiatry 1984;47:496-99.

DATA

John Major: 30 30 29 27 27 27 26 26 25 25 24 24 23 20
Not Margaret Thatcher: 29 22 20 19 18 14
Margaret Thatcher: 23 20 19 19 18 18 16 15 14 13 12 10 10
Other (e.g. Churchill, MacMillan): 14 14 14 12 12 8 8

Score



Note: I made this diagram in SYSTAT.

Comparison of several groups .. continued

```
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR  40 CASES
  DEPENDENT VARIABLE IS    SCORE
  GROUPING VARIABLE IS    GROUP

        GROUP  COUNT   RANK SUM   Ave Rank   Ave(SD) of Score

 JM         1    14      454         32.4         25.9(2.8)
 Not MT     2     6      133         22.2         20.2(5.0)
 MT         3    13      184         14.2         15.9(4.0)
 O          4     7       48          6.9         11.7(2.7)


KRUSKAL-WALLIS TEST STATISTIC =    28.15534
PROBABILITY IS     0.00000 ASSUMING CHI-SQUARE DISTRIBUTION
WITH  3 DF
```

Note from JH: I don't like the fact that the authors did so many "pairwise" comparisons of the 4 groups... It could be that there is a general downwards trend, but one doesn't have enough data in any one group to have the power to detect small differences between adjacent groups (in the limit, if we had groups spread out along a continuum on the x axis, we would not have enough data points at any one 'x' value). Thus, a single test of a trend is more relevant.

Here since, the answers have a logical ordering to them, a more sensitive test would be a **'trend' test** that takes account of this ordering.. in order to do this, we would need to put the 'distances' between the answers on some kind of interval scale, maybe 0, 1,2,3. We could then regress the score (or the rank corresponding to the score) on this 'distance' between answers and look at the magnitude of the slope.

The same logic applies to the test of trend in proportions, which we consider in §8.3 (illness in relation to number of times fell while wind surfing, and success in getting to play in professional football [soccer] in relation to when during the year one was born).

The last word on this...

Who's the PM?

SIR,—Having recently spent time in Africa and continental Europe I was aware of different perceptions of political arrangements in the UK and felt these might, in some settings, weaken the validity of the screening test discussed by Dr Fuller and Dr Meeran (June 1, p 1362). The recent gatherings of our extended family and friends, one in the UK and one abroad, afforded an opportunity to test this thesis. I conducted informal polls of those attenders at these gatherings in June, 1991. Of 24 people living principally in the UK 100% thought that Mr Major was Prime Minister. Of 29 living outside the UK 48% thought that Mrs Thatcher was Prime Minister, 10% named Mr Major, 20% thought it was "some man", and 22% "did not know". Of the UK residents 7 also volunteered that Mrs Thatcher thought she was still Prime Minister, and 2 of these added that Mr Major did not seem sure whether he or she was in charge. To my judgment all the participants would have performed well on the mini mental state test. My findings suggest that for the immediate future the "Who is the British Prime Minister?" test may be less than useful for visitors to the UK—or for UK residents of a perceptive nature.

MARY BRAHAM Oakhurst Drive, Wickford: Essex, UK