NON-PARAMETRIC TESTS:

What are non-parametric tests?

Statistical tests fall into two kinds: parametric tests assume that the data on which they are used possess certain characteristics, or "parameters". If the data do not possess these features, then the results of the test may be invalid. The parameters which are taken for granted are:

(a) the data are roughly normally distributed;

(b) the data show homogeneity of variance; and

(c) the data were measured on an interval or ratio scale.

If any or all of these assumptions are untrue, then it is safest to use a non-parametric test instead - a statistical test which does not make any of these assumptions about the nature of the data.

We have already encountered some of these tests. Here are examples of parametric tests and their non-parametric counterparts:

Parametric test:	**Non-parametric counterpart:**
Pearson correlation coefficient	Spearman's correlation coefficient
Independent-means t-test	Mann-Whitney test
Dependent-means t-test	Wilcoxon test
One-way Repeated-Measures Analysis of Variance (ANOVA)*	Friedman's test
One-way Independent Measures Analysis of Variance (ANOVA)*	Kruskal-Wallis test

***** ANOVA will be covered in the Second Year.

Chi-Square tests are another kind of non-parametric test, useful with frequency data (number of subjects falling into various categories). The tests dealt with in this handout are used when you have one or more *scores* from each subject. All four tests covered here - Mann-Whitney, Wilcoxon, Friedman's and Kruskall-Wallis - involve ranking the data, and then seeing how big the differences are between the sums of the ranks.

The Mann-Whitney test:

Use this when you have two conditions, and each condition is performed by a separate group of subjects. Each subject produces one score. The question the test answers is: is there a statistically significant difference between the two groups?

Step-by-step example:

Does it make any difference to students' comprehension of statistics whether the lectures are in English or in Serbo-Croat? We allocate students randomly to one of two groups. Each group is given a term's-worth of statistics lectures, followed by a statistics exam. The only difference between the two groups is that the lecturer speaks in Serbo-Croat to one group, and in English to the other group. Here are the results of the statistics exam:

English group (raw scores)	**English group (ranks)**	**Serbo-Croat group (raw scores)**	**Serbo-Croat group (ranks)**
18	17	**17**	15
15	10.5	**13**	8
17	15	**12**	5.5
13	8	**16**	12.5
11	3.5	**10**	1.5
16	12.5	**15**	10.5
10	1.5	**11**	3.5
17	15	**13**	8
		12	5.5

Step 1:

Rank all the scores together, ignoring for the moment which group they come from. The raw data are shown in bold in the table above, and the ranks are given in lighter type, in columns 2 and 4. (A brief revision of how to rank data is given at the end of this handout). Note that, as with an independent-means t-test, it is not necessary to have equal numbers of subjects in the two groups.

Step 2:

Add up the ranks for group 1, to get T1. Here, T1 = 83.
Add up the ranks for group 2, to get T2. Here, T2 = 70.

Step 3:

N1 is the number of subjects in group 1; N2 is the number of subjects in group 2. Here, N1 = 8 and N2 = 9.

Step 3:

Select the larger of these two rank totals, and call this Tx. Here, Tx = 83. Nx is the number of subjects in this group; in this case, Nx = 8.

Step 4:

Find U:

$$U = N1 * N2 + \frac{Nx\,(Nx + 1)}{2} - Tx$$

In our example,

$$U = 8 * 9 + \frac{8\,(8 + 1)}{2} - 83 \qquad = 72 + 36 - 83 = 25$$

(N.B.: if there are unequal numbers of subjects - as in the present case - calculate U for *both* rank totals and then use the *smaller* U. For T1 in the present example, U = 25, and for T2, U = 47; therefore, we would use 25 as our value of U).

Step 5:

Look up the critical value of U, in the appropriate table, taking into account N1 and N2. If our obtained U is *smaller* than the critical value of U, we reject the null hypothesis and conclude that our two groups do differ significantly.

In the present example, the critical value of U for N1 = 8 and N2 = 9 is 15. Our obtained U of 25 is *larger* than this, and so we conclude that there is *no* significant difference between our two groups. In other words, performance in the statistics exam is unaffected by whether the lectures are given in English or in Serbo-Croat.

The Wilcoxon test:

Use this when you have two conditions, and both conditions are performed by the *same* subjects. Each subject produces two scores, one for each condition. The question this test answers is: is there a statistically significant difference between the two conditions?

Step-by-step example:

Does background music affect the performance of factory workers? We take a group of eight workers, and measure each worker's productivity (in terms of the number of flangle-grommets manufactured per hour) twice -once while the worker is listening to background music, and once while the same worker is working in silence. (For half of the subjects, the two conditions are administered in the order "music then silence", while for the remaining subjects the two conditions are given in the opposite order). Here are the workers' scores:

Worker:	**No music**	**Music**	**difference**	**rank**
1	15	10	5	4.5
2	12	14	-2	2.5
3	11	11	0	ignore
4	16	11	5	4.5
5	14	4	10	6
6	13	1	12	7
7	11	12	-1	1
8	8	10	-2	2.5

Step 1:

Find the difference between each pair of scores, keeping track of the sign of the difference. Thus, for subject one, 15 - 10 = 5. For subject two, 12 - 14 = -2. The results of this step are shown in the column entitled **difference**, in the table above.

Step 2:

Rank the differences, ignoring their sign. Thus the smallest difference is -1, so this gets a rank of 1. The next smallest difference is -2, but there are two of these; therefore they get the average of ranks 2 and 3: 2.5. The results of this step are shown in the column entitled **rank**. Ignore any difference-scores of zero, which occur if a subject's pair of scores is identical.

Step 3:

Add together the ranks belonging to scores with a positive sign. Here, the positive-signed ranks sum to 22.

Add together the ranks belonging to scores with a negative sign. Here, the negative-signed ranks sum to 6.

Step 4:

"W" is the *smaller* sum of ranks; so here, W = 6.

N is the number of differences, omitting zero differences. Here, N = 8 - 1 = 7.

Step 5:

From a table, find the critical value of W, for your N. Your obtained W has to be *smaller* than this crtitical value, for it to be statistically significant.

In our example, the critical value of W for an N of 7, is 2. Our obtained W of 6 is bigger than this, and so we would conclude that there was no statistically significant difference between our two conditions. Worker productivity appears to be unaffected by the presence or absence of background music.

Kruskall-Wallis test:

This is similar to the Mann-Whitney test, except that it enables you to compare three or more groups rather than just two. *Different* subjects are used for each group.

Step by step example:

Does it make any difference to student comprehension of statistics whether the lectures are given in English, Serbo-Croat or Cantonese? We take three groups of psychology students, and give them similar lectures; the only difference is in the language used by the lecturer. One group gets lectures in English; one group gets lectures in Serbo-Croat; and the remaining group gets lectures in Cantonese. As with the Mann-Whitney example, the dependent variable is student exam performance.

Here are the data:

English (raw score)	English (rank)	Serbo-Croat (raw score)	Serbo-Croat (rank)	Cantonese (raw score)	Cantonese (rank)
20	3.5	**25**	7.5	**19**	1.5
27	9	**33**	10.5	**20**	3.5
19	1.5	**35**	10.5	**25**	7.5
23	6	**36**	12	**22**	5

Step 1:

Rank the scores, disregarding which group they belong to. The ranks are given in light type in the table above. The lowest score is 19, but there are two of these (one in the English group and the other in the Cantonese group); therefore their rank is the average of ranks 1 and 2 (= 1.5). The next lowest score is 20, and again there are two of these, so they get th average of ranks 3 and 4 (= 3.5).

Step 2:

Find the total of the ranks for each group separately. T1 (the total for the Enlgish group) is 20; T2 (for the Serbo-Croat group) is 40.5; and T3 (for the Cantonese group) is 17.5.

Step 3:

Find H.

$$\mathbf{H} = \left[\frac{12}{N(N+1)} * \sum \frac{Tc^2}{n_c} \right] - 3 * (N+1)$$

where N is the total number of subjects;
Tc is the rank total for each group;
and nc is the number of subjects in each group.

Here,

$$\sum \frac{Tc^2}{n_c} = \frac{20^2}{4} + \frac{40.5^2}{4} + \frac{17.5^2}{4} = 586.62$$

$$\mathbf{H} = \left[\left(\frac{12}{12 * 13} \right) * 586.62 \right] - 3 * 13 = 6.12$$

Step 4:

The degrees of freedom is the number of groups minus one.
So here, d.f. = 3 - 1 = 2.

Step 5:

In most cases, compare H to the critical Chi-Square value for the degrees of freedom that you have. H is statistically significant if it is *larger* than the critical value of Chi-Square. In our example, H is 6.12. This is larger than 5.99, the critical value of Chi-Square for 2 d.f. Therefore we would conclude that the three groups show a significant difference in their statistics exam performance; the language in which statistics is taught does make a difference.

Note that the Kruskal-Wallis test merely tells you that the three groups differ in some way; it does not tell you where the difference comes from. It might be that the three groups all differ from each other, or it might be that two groups are similar to each other but both are different from the third. Inspection of the group medians can normally help you to decide what's going on.

Friedman's Test:

This is similar to the Wilcoxon test, except that you can use it with three or more conditions. Each subject participates in *all* of the different conditions of the experiment.

Step by step example:

Suppose we did our worker productivity experiment again. This time, instead of merely comparing the effects on productivity of music and silence, we have three conditions: silence, "easy-listening" music and marching-band music. We take five workers, and measure each worker's productivity three times (once while working under each type of background music). Note that, to avoid practice and fatigue effects, the order of presentation of each of these conditions should be varied. Here are the data.

	No Music (raw score)	**No Music (ranked score)**	**Easy listening (raw score)**	**Easy listening (ranked score)**	**Marching band (raw score)**	**Marching band (ranked score)**
Worker 1:	**4**	1	**5**	2	**6**	3
Worker 2:	**2**	1	**7**	2.5	**7**	2.5
Worker 3:	**6**	1.5	**6**	1.5	**8**	3
Worker 4:	**3**	1	**7**	3	**5**	2
Worker 5:	**3**	1	**8**	2	**9**	3

Step 1:

Rank *each subject's scores* individually. Thus worker 1 gives scores of 4, 5 and 6 for the English, Serbo-Croat and Cantonese conditions, so these get ranks of 1, 2 and 3 respectively. Worker 4's scores are 3, 7 and 5 for the three conditions, and so these scores get ranks of 1, 3 and 2 respectively.

Step 2:

Find the rank total for each condition, using the ranks from all subjects within that condition. Thus the rank total for the "English" condition above is 1+1+1.5+1+1 = 5.5. The rank total for the Serbo-Croat condition is 11. The rank total for the Cantonese condition is 13.5.

Step 3:

Work out χr^2 :

$$\chi \mathbf{r}^2 = \left[\left(\frac{\mathbf{12}}{\mathbf{N} * \mathbf{C} * (\mathbf{C}+\mathbf{1})}\right) * \sum \mathbf{Tc}^2\right] - \mathbf{3} * \mathbf{N} * (\mathbf{C}+\mathbf{1})$$

where C is the number of conditions,
N is the number of subjects,
and ΣTc^2 is the sum of the squared rank totals for each condition.

In other words, to get ΣTc^2,
(a) take each rank total and square it.
Here, $5.5^2 = 30.25$. $11^2 = 121$. $13.5^2 = 182.25$.
(b) Add these squared totals together. 30.25 + 121 + 182.25 = 333.5.

In our example,

$$\chi \mathbf{r}^2 = \left[\left(\frac{\mathbf{12}}{\mathbf{5} * \mathbf{3} * \mathbf{4}}\right) * \mathbf{333.5}\right] - \mathbf{3} * \mathbf{5} * \mathbf{4} \quad = \quad \mathbf{6.7}$$

Step 4:

The degrees of freedom are given by the number of conditions minus one. Here, C - 1 = 3 - 1 = 2.

Step 5:

Assessing the statistical significance of Xr^2 depends on the number of subjects and the number of groups. If you have more than 9 subjects, you can use a Chi-Square table. (If you have less than nine subjects, special tables of critical values are available - for example tables C(1) or C(11) in Greene and D'Oliveira's book). Compare your obtained Xr^2 value to the critical value of Chi-Square for your d.f. If your obtained value of Xr^2 is *bigger* than the critical Chi-Square value, the difference between your conditions is statistically significant. As with the Kruskal-Wallis test, Friedman's test only tells you that some kind of difference exists; inspection of the median score for each condition will usually be enough to tell you exactly where the difference comes from.

Revision of how to Rank the data:

We rank the data in exactly the same way as we did for the Spearman's correlation coefficient last term.

(a) Rank the data separately for each variable. Assign a rank of “1” to the lowest score; “2” to the next lowest; and so on.

(b) If two or more scores have the same value, they are said to be “tied”. In this case, you need to go through the following procedure.

(i) Start by giving each tied score the rank which it would have been given, had it been different from the other scores.

(ii) Add together the ranks for the tied scores, and divide by the number of tied scores. This gives you an average rank. Assign this average rank to each of the tied scores.

(iii) Give the next score after the set of tied scores, the rank which it would have obtained had the scores not been tied.

All this is a lot more complicated to explain than to do. Hopefully, the following examples make the procedure clear:

raw score: 12 15 15 16

Here, we have two ties, since two scores have the same value (15).
Had there been no ties, the ranks would have been as follows:

“original” rank: 1 2 3 4

Instead, add together the ranks for the tied scores and divide by the number of tied scores. (2+3)/2 = 2.5. Each tied score gets this average rank of 2.5:

“actual” rank: 1 2.5 2.5 4

Note that the raw score of 16 (the first score to be encountered after the set of ties) gets the rank of 4 that it would have obtained if no ties had been present.(We have effectively “used up” the ranks of 2 and 3 in dealing with the scores of 15 and 15).