The Mann-Whitney test:

Use this when two different groups of participants perform *both* conditions of your study: i.e., it is appropriate for analysing the data from an independent-measures design with two conditions. Use it when the data do not meet the requirements for a parametric test (i.e. if the data are not normally distributed; if the variances for the two conditions are markedly different; or if the data are measurements on an ordinal scale). Otherwise, if the data meet the requirements for a parametric test, it is better to use an independent-measures *t*-test (also known as a "two-sample" *t*-test).

The logic behind the Mann-Whitney test is to rank the data for each condition, and then see how different the two rank totals are. If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one. As a result, the rank totals will be quite different. On the other hand, if the two conditions are similar, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar. The Mann-Whitney test statistic "*U*" reflects the difference between the two rank totals. The SMALLER it is (taking into account how many participants you have in each group) then the less likely it is to have occurred by chance. A table of critical values of *U* shows you how likely it is to obtain your particular value of *U* purely by chance. Note that the Mann-Whitney test is unusual in this respect: normally, the BIGGER the test statistic, the less likely it is to have occurred by chance).

This handout deals with using the Mann-Whitney test with small sample sizes. If you have a large number of participants, you can convert *U* into a *z*-score and look this up instead. The same is true for the Wilcoxon test. There is a handout on my website that explains how to do this, for both tests.

Step by step example of the Mann-Whitney test:

Suppose we want to compare the effectiveness of two laxatives ("Flushem" and "Kleerout"). We find twelve constipated volunteers and randomly allocate half of them to use Flushem, and the rest to use Kleerout. We then get them to rate the effectiveness of their laxative, on a ten-point scale (where "0" = "very ineffective" and "10" = "very effective"). Here are their ratings:

Flushem		**Kleerout**	
Participant	**Rating**	**Participant**	**Rating**
1	**3**	**1**	**9**
2	**4**	**2**	**7**
3	**2**	**3**	**5**
4	**6**	**4**	**10**
5	**2**	**5**	**6**
6	**5**	**6**	**8**
median:	**3.5**		**7.5**

Is there a significant difference in the rated effectiveness of the two laxatives? The medians for the two conditions do look different (it looks as if Kleerout might be rated as more effective), but the Mann-Whitney test will confirm whether this difference is large enough to be statistically significant (i.e. unlikely to have occurred by chance).

We have two conditions, with each participant taking part in only one of the conditions. The data are ratings (ordinal data), which is why we are using the nonparametric Mann-Whitney test, rather than an independent measures *t*-test.

Step 1:

Rank all scores together, ignoring which group they belong to. The lowest score gets a rank of "1", the next lowest gets a rank of "2", and so on. If two or more scores are identical, this is a "tie". They get the average of the ranks that they would have obtained, had they been different from each other. Here, we have two scores with a value of 2. they therefore get the average of the ranks "1" and "2", which is 3/2 = 1.5. This procedure "uses up" the ranks of "1" and "2", so the next highest score (3) gets a rank of "3".

Flushem			**Kleerout**		
Participant	**Rating**	**Rank**	**Participant**	**Rating**	**Rank**
1	**3**	**3**	**1**	**9**	**11**
2	**4**	**4**	**2**	**7**	**9**
3	**2**	**1.5**	**3**	**5**	**5.5**
4	**6**	**7.5**	**4**	**10**	**12**
5	**2**	**1.5**	**5**	**6**	**7.5**
6	**5**	**5.5**	**6**	**8**	**10**

Step 2:

Add up the ranks for Flushem, to get T1.

T1 = 3 + 4 + 1.5 + 7.5 + 1.5 + 5.5 = 23

Step 3:

Add up the ranks for Kleerout, to get T2.

T2 = 11 + 9 + 5.5 + 12 + 7.5 + 10 = 55

Step 4:

Select the larger of these two rank totals, and call it TX. In this case TX is the rank total for Kleerout, which is 55.

Step 5:

Calculate N1, N2 and NX.

N1 is the number of people in the group that gave you the T1 rank total (Flushem users); N2 is the number of people in the group that gave you the T2 rank total (Kleerout users); and NX is the number of people in the group that gave the larger rank total, TX (in this case, the number of people in the Kleerout group). In our example, these numbers are all the same (6) because we have equal numbers of participants in our two groups. However, this isn't necessarily the case, and so the Mann-Whitney formula takes account of this by getting you to enter N1, N2 and NX separately.

Step 6:

Find *U* by working through the formula below. Remember that Tx is the larger rank total.

$$U = N1 * N2 + NX * \frac{(NX+1)}{2} - TX$$

$$U = 6 \times 6 + 6 \times \frac{(6+1)}{2} - 55$$

$$U = 6 \times 6 + 6 \times \frac{(7)}{2} - 55$$

U = 36 + 21 - 55

U = 57 - 55

U = 2

Step 7:

Use a table of critical *U* values for the Mann-Whitney test (such as the one on my website, reproduced below).

N_1 \ N_2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

This table shows critical values of *U* for different group sizes, for a two-tailed test at the .05 significance level. (Tables also exist for higher significance levels, such as .01 and .001, but only the .05 table is shown on my website). For N1 = 6 and N2 = 6, the critical value of *U* is 5. To be statistically significant, our obtained *U* has to be equal to or LESS than this critical value. (Note that this is different from many statistical tests, where the obtained value has to be equal to or *larger* than the critical value).

Our obtained *U* = 2, which is less than 5. Therefore our obtained value of *U* is even less likely to occur by chance than the one in the table: we can conclude that the difference that we have found between the ratings for the two laxatives is unlikely to have occurred by chance. It looks as if participants' assessments of the laxative's effectiveness do indeed differ.

(Inspection of the medians suggests that Kleerout is rated as being more effective than Flushem. However we initially predicted only that there would be *some kind of difference* between the two laxatives, and not that Kleerout would be better than Flushem: we have therefore conducted a non-directional, two-tailed, test, and strictly speaking all we can conclude from it is that the two laxatives differ in rated effectiveness).