The Kruskal-Wallis test:

This test is appropriate for use under the following circumstances:
(a) you have three or more conditions that you want to compare;
(b) each condition is performed by a *different* group of participants; i.e. you have an independent-measures design with three or more conditions.
(c) the data do not meet the requirements for a parametric test. (i.e. use it if the data are not normally distributed; if the variances for the different conditions are markedly different; or if the data are measurements on an ordinal scale).

If the data meet the requirements for a parametric test, it is better to use a one-way independent-measures Analysis of Variance (ANOVA) because it is more powerful than the Kruskal-Wallis test.

Step by step example of the Kruskal-Wallis test:

Does physical exercise alleviate depression? We find some depressed people and check that they are all equivalently depressed to begin with. Then we allocate each person randomly to one of three groups: no exercise; 20 minutes of jogging per day; or 60 minutes of jogging per day. At the end of a month, we ask each participant to rate how depressed they now feel, on a Likert scale that runs from 1 ("totally miserable") through to 100 (ecstatically happy").

The appropriate test here is the Kruskal-Wallis test. We have three separate groups of participants, each of whom gives us a single score on a rating scale. Ratings are examples of an *ordinal* scale of measurement, and so the data are not suitable for a parametric test.

The Kruskal-Wallis test will tell us if the differences between the groups are so large that they are unlikely to have occurred by chance. Here are the data:

Rating on depression scale:

	No exercise	**Jogging for 20 minutes**	**Jogging for 60 minutes**
	23	22	59
	26	27	66
	51	39	38
	49	29	49
	58	46	56
	37	48	60
	29	49	56
	44	65	62
mean rating (SD):	**39.63 (12.85)**	**40.63 (14.23)**	**55.75 (8.73)**

Step 1: Rank all of the scores, ignoring which group they belong to. The procedure for ranking is as follows: the lowest score gets the lowest rank. If two or more scores are the same then they are "tied". "Tied" scores get the average of the ranks that they *would* have obtained, had they not been tied. Here's the scores again, now with their ranks in brackets:

	C1 (No exercise)	**C2 (Jogging for 20 minutes)**	**C3 (Jogging for 60 minutes)**
	23 (2)	22 (1)	59 (20)
	26 (3)	27 (4)	66 (24)
	51 (16)	39 (9)	38 (8)
	49 (14)	29 (5.5)	49 (14)
	58 (19)	46 (11)	56 (17.5)
	37 (7)	48 (12)	60 (21)
	29 (5.5)	49 (14)	56 (17.5)
	44 (10)	65 (23)	62 (22)
mean rank (SD)	**9.56 (6.25)**	**9.94 (6.84)**	**18.00 (5.09)**
sum of ranks (Tc)	**76.5**	**79.5**	**144**

In detail, this is how the ranks are arrived at for these scores.

(a) "22" is the lowest score. This gets a rank of 1.

(b) "23" is the next lowest score. This gets a rank of 2.

(c) "26" is the next lowest score. This gets a rank of 3.

(d) "27" is the next lowest score. This gets a rank of 4.

(e) There are two instances of "29". This is a "tie". They both get the average of the ranks that they would have been allocated, had they been different from each other. So the next two ranks are 5 and 6. The average of 5 and 6 is 11/2 = 5.5. Both instances of "29" therefore get a rank of 5.5.

(f) "37" is the next lowest score. This gets a rank of 7 (because we've just "used up" ranks 5 and 6).

(g) "38" is the next lowest score, and it gets a rank of 8.

(h) "39" is the next lowest score, and it gets a rank of 9.

(i) "44" gets a rank of 10, "46" gets a rank of 11, and "48" gets a rank of 12.

(j) There are three instances of "49", so this is another tie. They each get the average of the next three unused ranks ((13+14+15) / 3 = 14).

(k) "51" is the next lowest score, and it gets the next "unused" rank, which is 16.

(l) There are two instances of "56", so they get the average of the next two unused ranks ((17+18) /2 = 17.5).

(m) "58" gets the next unused rank, which is 19.

(n) "59" gets a rank of 20, "60" gets 21, "62" gets 22, "65" gets 23, and 66 gets 24.

This is all tedious, but really not difficult to do once you've practiced it a couple of times!

Step 2: Find "Tc", the total of the ranks for each group. Just add together all of the ranks for each group in turn.

Here, Tc1 (the rank total for the "no exercise" group) is 76.5.

Tc2 (the rank total for the "20 minutes" group) is 79.5.

Tc3 (the rank total for the "60 minutes" group) is 144.

Step 3: Find "H".

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

N is the total number of participants (all groups combined). We have 24 participants (3 groups of 8).

Tc is the rank total for each group. Tc1 = 76.5, Tc2 = 79.5, and Tc3 = 144.

nc is the number of participants in each group. Here, nc1 = 8, nc2 = 8 and nc3 = 8.

For our data,

$$H = \left[\frac{12}{24*(24+1)} * \sum \frac{Tc^2}{n_c} \right] - 3 * (24+1)$$

$\sum \frac{Tc^2}{n_c}$ means the following:

First, take each group's rank total, square it and then divide the result by the number of participants in that group.

Then, add these numbers together.

$$\sum \frac{Tc^2}{n_c} = \frac{76.5^2}{8} + \frac{79.5^2}{8} + \frac{144^2}{8}$$

$$\sum \frac{Tc^2}{n_c} = 731.5313 + 790.0313 + 2592.0000 = \mathbf{4113.5625}$$

$$H = \left[\frac{12}{600} * 4113.5625\right] - 75$$

$$H = \left[0.02 * 4113.5625\right] - 75$$

$$H = \left[0.02 * 4113.5625\right] - 75$$

$$H = \mathbf{7.27}$$

Step 4: the degrees of freedom is the number of groups minus one. Here we have three groups, and so we have 2 d.f.

Step 5:

Assessing the significance of *H* depends on the number of participants and the number of groups.

If you have three groups, with five or fewer participants in each group, then you need to use the special table for small sample sizes (which is on my website).

If you have more than five participants per group, then treat *H* as Chi-Square. H is statistically significant if it is equal to or larger than the critical value of Chi-Square for your particular d.f. (The table of Chi-Square values is also on my website).

Here, we have eight participants per group, and so we treat *H* as Chi-Square. *H* is 7.27, with 2 d.f. Here's the relevant part of the Chi-Square table:

Table of critical Chi-Square values:

df	*p* = .05	*p* = .01	*p* = .001
1	3.84	6.64	10.83
2	**5.99**	**9.21**	**13.82**
3	7.82	11.35	16.27

Look along the row that corresponds to your number of degrees of freedom.

So in this case, we look along the row for 2 d.f.

We compare our obtained value of *H* to each of the critical values in that row of the table, starting on the lefthand side and stopping once our value of *H* is no longer equal to or larger than the critical value.

So here, we start by comparing our *H* of 7.27 to 5.99. With 2 degrees of freedom, a value of Chi-Square as large as 5.99 is likely to occur by chance only 5 times in a hundred: i.e. it has a *p* of .05. Our obtained value of 7.27 is even larger than this, and so this tells us that our value of *H* is even *less* likely to occur by chance. Our *H* will occur by chance with a probability of *less* than 0.05.

Move on, and compare our *H* to the next value in the row, 9.21. 9.21 will occur by chance one time in a hundred, i.e. with a *p* of .01. However, our *H* of 7.27 is less than 9.21, not bigger than it. This tells us that our value of H is not so large that it is likely to occur with a probability of 0.01.

Conclusion:

The likelihood of obtaining a value of *H* as large as the one we've found, purely by chance, is somewhere between 0.05 and 0.01 - i.e. pretty unlikely, and so we would conclude that there is a difference of some kind between our three groups.

Note that the Kruskal-Wallis test merely tells you that the groups *differ* in some way: you need to inspect the group means or medians to decide precisely how they differ. However in this particular case, the interpretation seems fairly straightforward: exercise does seem to reduce self-reported ratings of depression, but only in the case of participants who are doing an hour of it. There seems to be no difference between those participants who took 20 minutes of exercise per day, and those who did not exercise at all.

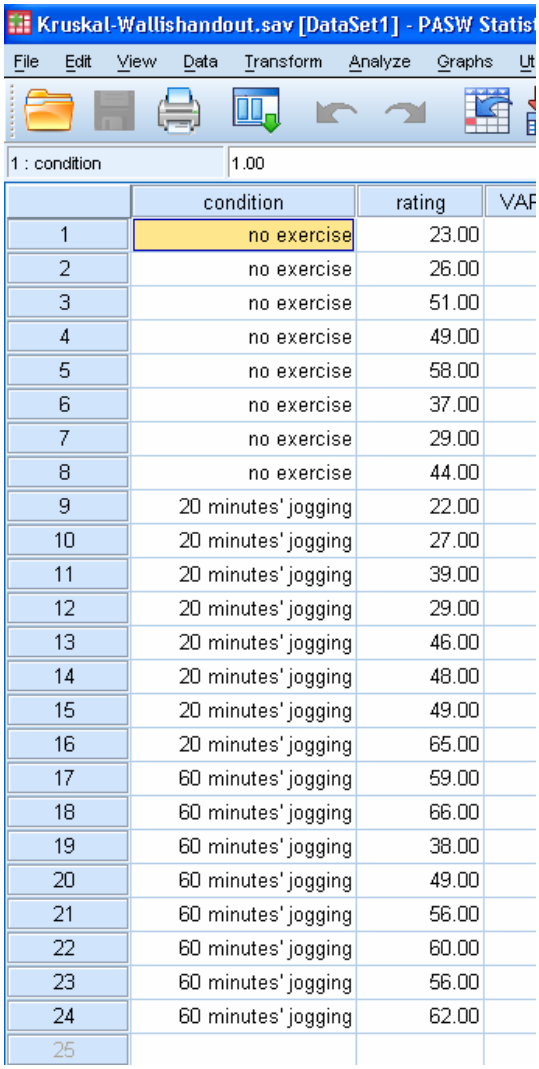
We could write this up as follows:

"A Kruskal-Wallis test revealed that there was a significant effect of exercise on depression levels (H (2) = 7.27, $p < .05$). Inspection of the group means suggests that compared to the "no exercise" control condition, depression was significantly reduced by 60 minutes of daily exercise, but not by 20 minutes of exercise". (NB: note that a higher score in this study equates to a higher level of mood and hence a *lower* level of depression).

Using SPSS to perform the Kruskal-Wallis test:

Step 1:

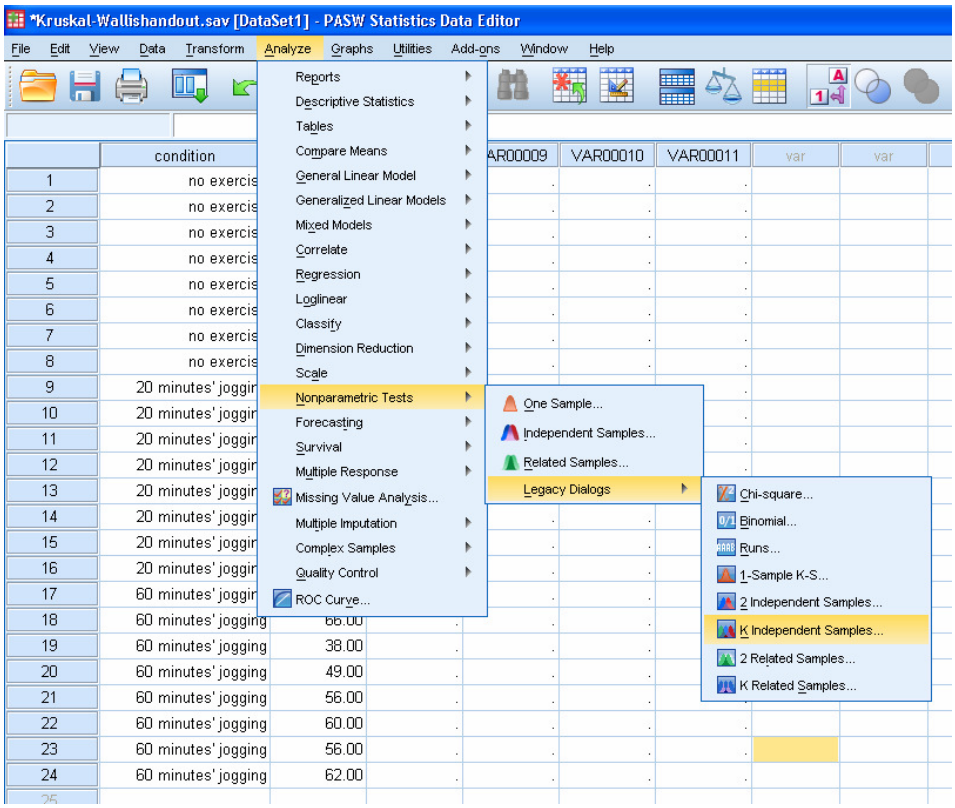
Enter the data into SPSS. This is an independent-measures design, so you need two columns. One (labelled "condition" here) tells SPSS which condition each participant was in. I used the codes "1", "2" and "3" for "no exercise", "20 minutes' jogging" and "60 minutes' jogging" respectively. I then changed to "variable view" and gave the codes "value labels", to make it easier to see which condition was which. The second column ("rating") gives the corresponding scores. Thus, in effect, each row in the spreadsheet corresponds to a single participant - it tells SPSS which condition that person was in, and what their depression rating was.



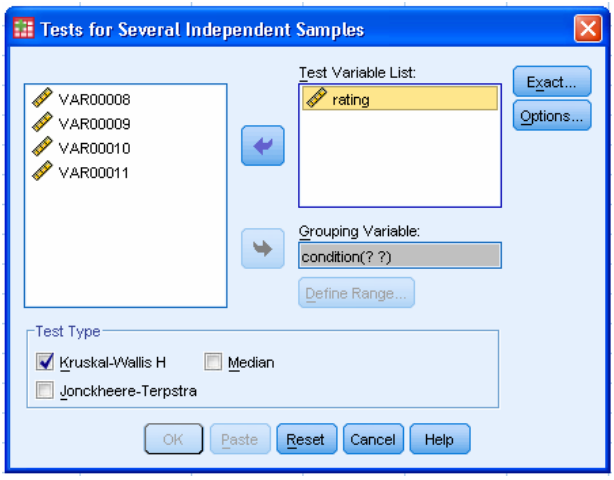
	condition	rating	VAR
1	no exercise	23.00	
2	no exercise	26.00	
3	no exercise	51.00	
4	no exercise	49.00	
5	no exercise	58.00	
6	no exercise	37.00	
7	no exercise	29.00	
8	no exercise	44.00	
9	20 minutes' jogging	22.00	
10	20 minutes' jogging	27.00	
11	20 minutes' jogging	39.00	
12	20 minutes' jogging	29.00	
13	20 minutes' jogging	46.00	
14	20 minutes' jogging	48.00	
15	20 minutes' jogging	49.00	
16	20 minutes' jogging	65.00	
17	60 minutes' jogging	59.00	
18	60 minutes' jogging	66.00	
19	60 minutes' jogging	38.00	
20	60 minutes' jogging	49.00	
21	60 minutes' jogging	56.00	
22	60 minutes' jogging	60.00	
23	60 minutes' jogging	56.00	
24	60 minutes' jogging	62.00	
25			

Step 2:

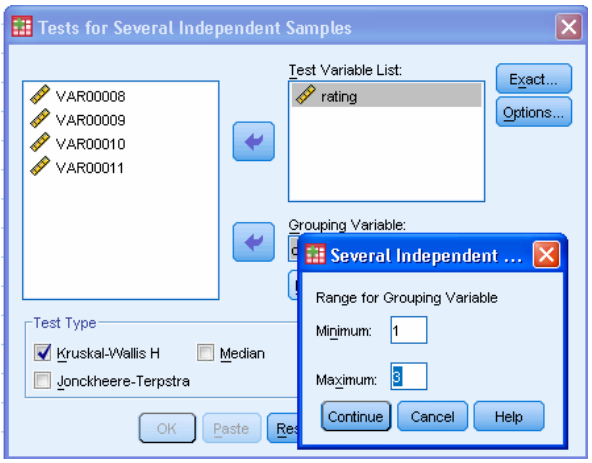
On the top menu, select ***Analyze,*** pick ***nonparametric tests.*** In SPSS version 18 (the one I'm using) you then choose ***legacy dialogs***, and finally ***k independent samples...*** (In earlier versions of SPSS, this sequence simply goes as follows: ***Analyze***, ***nonparametric tests***, ***k independent samples...***).



The following dialogue box appears:



You move the variable containing your scores ("rating" in this case) into the box labelled "test variable list". You move the variable containing the codes that identify the different conditions ("condition" in this case) into the box labelled "grouping variable". Then click on "define range" and tell SPSS about the codes for the various conditions, as follows:



I have three conditions, so the minimum is "1" and the maximum is "3". Then click on "continue". Next, click on options and select "descriptive statistics" to get the mean for each condition. Finally, click on "OK", to perform the test.

Here's what the output looks like:

Ranks

	depression rating (1 = low, 7 = high)	N	Mean Rank
rating	no exercise	8	9.56
	20 minutes' jogging	8	9.94
	60 minutes' jogging	8	18.00
	Total	24	

Test Statistics[a,b]

	rating
Chi-square	7.290
df	2
Asymp. Sig.	.026

a. Kruskal Wallis Test

b. Grouping Variable: depression rating (1 = low, 7 = high)

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
rating	24	45.3333	13.85222	22.00	66.00
depression rating (1 = low, 7 = high)	24	2.0000	.83406	1.00	3.00

The first box tells you what the dependent variable was ("depression rating" in this case); what the names of the conditions were; "N", the number of participants in each condition; and the mean rank for each condition (not particularly useful).

The second box gives you the result of the Kruskal-Wallis test as a value of Chi-Square ; how many d.f. are associated with it; and the significance level (an exact p-value, as opposed to the approximate value that we have to use if we do the test by hand and use a table to look up its probability value).

The third box gives you the descriptive statistics. Note that SPSS has unintelligently calculated the mean and standard deviation for both the grouping variable (condition) *and* the dependent variable (rating). The descriptive statistics for the grouping variable should be ignored, as they are quite meaningless.

Note that the value of *H* is not quite the same as the one we worked out by hand: SPSS says it's 7.29, whereas by hand it came to 7.27. By hand, we estimated that the probability of obtaining a value of *H* this large by chance would be somewhere between .05 and .01. SPSS enables us to be a bit more precise, estimating the probability to be .026 (.03, if you report it to 2 significant digits).