Inferential statistical tests for experiments

This section provides a description and explanation of inferential statistical analysis for experiments. This will help you to understand:

1 The purpose of inferential statistical analysis in the scientific process

2 Which tests to choose for different experimental research settings

3 The procedures involved for carrying out individual tests.

Descriptive statistical analysis does exactly what it suggests; it describes the findings in numerical (table), visual (graph) and verbal (words) forms (see 'Descriptive statistical analysis', page 37). However, inferential statistical analysis is more complex and sophisticated. It involves the use of statistical tests to assess, in the case of experiments, whether measured differences between two conditions of an independent variable are *significant differences* (beyond the boundaries of chance) that can be generalized from the sample tested to the whole target population that the sample represents.

If a coin is tossed 100 times, then by the law of averages there should be 50 heads and 50 tails. However, it might be 52 heads and 48 tails, which means there is a difference between the two sets of data (52 and 48), but is it beyond the boundaries of chance for that to occur? Would you suspect something was amiss, like the coin being weighted on one side or that the person tossing the coin was manipulating it in some way? Probably not. But where would you 'draw the line' between a result being within chance factors and falling outside them? 60 heads and 40 tails? 90 heads and 10 tails? This is what inferential statistical tests do, they use the concept of *probability*, through the use of statistical calculations, to set a 'cut-off point' that determines if differences between sets of experimental data are 'real' (significant) differences beyond the boundaries of chance.

Probability

Probability concerns how certain researchers are that a difference between two sets of data is a 'real' (significant) difference. There is no such thing as absolute certainty (100% probability) that a difference was not due to chance factors. There is a slight possibility that debris from space, hurtling through the atmosphere, will strike you dead in the next few seconds; it probably won't happen, but there is a slight possibility that it could happen.

This is why in science it is impossible to 'prove' something beyond all doubt. Therefore, we need a 'cut-off' point' beyond which we will accept a statistical test result as showing a significant difference.

In psychology, a significance (probability) level of 5% is used, which is expressed as:

$p \leq 0.05$

This means that there is a 5% possibility that an observed difference between two sets of data (such as that between a coin landing heads or tails when tossed), which is said by an inferential statistical test to be 'significant' (beyond the boundaries of chance), is actually due to chance factors. This means the null hypothesis would be wrongly rejected and the experimental hypothesis wrongly accepted. This phenomenon is known as a *Type I error*, for example when a pregnancy test says a woman is pregnant and she isn't. This is regarded in psychology as being an acceptable level of error. With a 5% significance level, it means that, on average, for every 100 significant differences found by inferential statistical tests, where the null hypothesis would be rejected, 5 of them would be wrong.

When experimenting in new research areas that haven't been explored before (or when testing potentially harmful stimuli, like newly manufactured drugs), it is customary to use a stricter, higher level of significance, such as a 1% significance level, which is expressed as:

$p \leq 0.01$

This means there is only a 1% possibility that an observed difference between two sets of data, said by an inferential test to be 'significant' is actually due to chance factors. An even stricter level of significance of $p \leq 0.001$ would mean there's a 99.9% certainty of an observed difference between two sets of data being beyond chance factors, but there would still be a 0.1% chance that the difference had occurred by chance.

There is a possibility when using strict levels of significance (such as a 1% level) that no significant difference will be found by an inferential statistical test, when in fact the observed difference was actually beyond chance factors, such as a pregnancy test saying a woman isn't pregnant when in fact she is. This would mean that the null hypothesis would be wrongly accepted and the experimental hypothesis wrongly rejected. This is called a *Type II error*. The stricter the level of significance used, the more chance there is of a Type II error occurring.

The reason a 5% significance level is used as the 'accepted level' in psychological research, is because it is seen as striking a balance between making a Type I and a Type II error.

Interpretation of significance

Inferential statistical tests produce an *observed value*, which is then compared to a critical value (*cv*) in a critical value table in order to determine if the observed value is significant and thus whether hypotheses can be accepted or rejected. What an actual *cv* will be depends upon whether an experimental hypothesis is one- or two-tailed (also known as directional and non-directional), the number of participants or participant pairs (*N*) used and what level of significance (usually the 5% level) is used.

The actual interpretations of observed and critical values of specific statistical tests will be referenced when detailing individual tests.

Choosing specific inferential statistical tests for experimental studies

Once an experiment has been carried out and data generated, an appropriate inferential statistical test must be selected. Choice of test, other than the fact that to analyse data generated from experiments requires a test of difference, is dependent on two factors:

1. *The experimental design used* – whether an independent group design (each participant only does one condition of the experiment) or repeated measures design (each participant does all conditions of the experiment) was used (a matched participants design is regarded as a type of repeated measures design).
2. *The level of data generated* – whether the data is of **nominal, ordinal** or **interval/ratio** level.

Key definitions

Nominal data – a crude, relatively uninformative level of data that involves frequencies, e.g. how many people prefer orange juice or lemonade.

Ordinal data – a more informative level of data that involves data which are rankable (can be put into rank order), e.g. the finishers in a running race. Data that is ordinal is also of nominal level.

Interval/ratio data – the most informative level of data that involves data of equal measurement intervals, e.g. seconds in time. Interval data has an arbitrary zero point, whereas ratio data has an absolute zero point. For instance, temperature is interval as there can be a minus reading (e.g. minus 15 degrees centigrade), while someone with zero pounds in their bank account has no money (ratio data). Data that is interval/ratio is also of nominal and ordinal level.

Once the experimental design and data levels have been determined, consult Table 5.1 for choice of appropriate test.

	Independent groups design	Repeated measures design
Nominal data	*Chi-squared test*	*Sign test*
Ordinal data	*Mann-Whitney test*	*Wilcoxon signed-matched ranks test*
Interval/ratio data	*Independent t-test*	*Repeated t-test*

Table 5.1 Inferential statistical tests

Tests based on nominal level data (chi-squared and sign test) are less sensitive (due to the uninformative nature of the data), which means that these tests are less able than more sensitive tests to detect a significant difference if there is one (and therefore make it easier to make a Type II error and wrongly accept a null hypothesis).

Tests based on ordinal level data (Mann-Whitney and Wilcoxon signed-matched ranks) are more sensitive (due to the more informative nature of the data) and so are more able to detect a significant difference if there is one (and thus less likely to lead to a Type II error and a null hypothesis being wrongly accepted). However, tests based on interval/ratio data (independent *t*-test and repeated *t*-test) are the most sensitive and thus most able to detect a significant difference if there is one (and thus least likely to lead to a Type II error and a null hypothesis being wrongly accepted).

The sign test

Criteria for choice

The sign test is used when a repeated measures design (RMD) has been used and the data is of at least nominal level.

Rationale of the test

The test compares the number of scores that go in one direction (e.g. prefer orange juice) to the number of scores that go in another direction (e.g. prefer lemonade) to see if any difference in direction of scores is beyond chance factors (is 'significant').

How to calculate the sign test

- Put the data into appropriate table form (see 'Make-believe example of sign test for blindfolded object-holding study', page 34).
- A plus (+) sign is put next to scores that go in one direction and a minus sign (–) to scores that go in the other direction (see Make-believe example).
- *s* = the number of times the less frequent sign occurs.
- Find the *cv* from an appropriate critical value table – this will be dependent on whether a one- or two-tailed hypothesis (directional or non-directional) has been used and the number of pairs of data used (*N*).
- If *s* is less than or equal to the *cv*, reject the null hypothesis.

Level of significance for a two-tailed test				
	0.05	0.025	0.01	0.005
Level of significance for a one-tailed test				
N	**0.10**	**0.05**	**0.02**	**0.01**
5	0	–	–	–
6	0	0	–	–
7	0	0	0	–
8	1	0	0	0
9	1	1	0	0
10	1	1	0	0
11	2	1	1	0
12	2	2	1	1
13	3	2	1	1
14	3	2	2	1
15	3	3	2	2
16	4	3	2	2
17	4	4	3	2
18	5	4	3	3
19	5	4	4	3
20	5	5	4	3

Table 5.2 Levels of significance for one- and two-tailed tests

The chi-squared test

Criteria for choice

The chi-squared test is used when an independent groups design (IGD) has been used and the data is of at least nominal level.

Rationale of the test

The test compares the expected frequencies of scores to the actual observed frequencies to see if they differ beyond chance factors. For example, comparing how many people accurately recalled a written statement about Paris with how many people accurately recalled a written statement about Glasgow (see page 28). If there is no significant difference between the frequency of the scores, then the expected frequencies should be similar to the observed frequencies.

How to calculate the chi-squared test

- Put the observed scores into a 2 × 2 contingency table, with the four cells referred to as A, B, C and D (see Table 4.4, page 31).

	Condition one	Condition two	Row total
First frequency	Cell A	Cell C	
Second frequency	Cell B	Cell D	
Column total			*GT*

Table 5.3 Chi-squared contingency table

- Calculate the expected frequencies (E) for each cell (A, B, C and D) from:

  E = row total × column total ÷ grand total

- The chi-squared (χ^2) statistic is then calculated from the expected frequencies (E) and observed frequencies (O) using the formula:

  $\chi^2 = \Sigma(O - E - \frac{1}{2})^2 \div E$

- See calculations for chi-squared test for the Bruner and Postman (1947) mini-practical, page 30.
- Find the *cv* from an appropriate critical value table – this will be dependent on whether you have a one- or two-tailed hypothesis (directional or non-directional), the level of significance used and the degrees of freedom (d.f.) (which with a 2 × 2 contingency table is 1).
- If χ^2 is equal to or greater than the *cv*, it is significant and so reject the null hypothesis.

d.f.	0.20	0.10	0.05	0.02	0.01	0.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.6	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.8	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.0	28.26	30.58	37.7
16	20.46	23.54	26.3	29.63	32.0	39.29
17	21.62	24.77	27.59	31.0	33.41	40.75
18	22.76	25.99	28.87	32.35	34.8	42.31
19	23.9	27.2	30.14	33.69	36.19	43.82
20	25.04	28.41	31.41	35.02	37.57	45.32
21	26.17	29.62	32.67	36.34	38.93	46.8
22	27.3	30.81	33.92	37.66	40.29	48.27
23	28.43	32.01	35.17	38.97	41.64	49.73
24	29.55	33.2	36.42	40.27	42.98	51.18
25	30.68	34.38	37.65	41.57	44.31	52.62
26	31.8	35.56	38.88	42.86	45.64	54.05
27	32.91	36.74	40.11	44.14	46.96	55.48
28	34.03	37.92	41.34	45.42	48.28	56.89
29	35.14	39.09	42.69	46.69	49.59	58.3
30	36.25	40.26	43.77	47.96	50.89	59.7

Table 5.4 Critical values of chi-squared for a two-tailed (non-directional) test. Chi-squared is significant if it is equal to or greater than the table value

d.f.	0.10	0.05	0.025	0.01	0.005	0.0005
1	1.64	2.71	3.84	5.41	6.64	10.83

Table 5.5 Critical values of chi-squared for a one-tailed (directional) test

The Mann-Whitney test

Criteria for choice

The Mann-Whitney test is used when an independent groups design has been used and the data is of at least ordinal level.

Rationale of the test

The test compares the ranks of two sets of data to see if they differ significantly from each other. If they do not, then the totals of the ranks of the two sets of data should be similar. For example, ranking the scores of people who sat a test in the room where they learned the information against people who sat the test in a different room, and then seeing if the total of the ranks for one condition is significantly different from that of the other condition (see page 34).

How to calculate the Mann-Whitney test

- Construct an appropriate table for calculating a Mann-Whitney test.

Participants Condition A:	Score	Rank	Participants Condition B:	Score	Rank
1			11		
2			12		
3			13		
4			14		
5			15		
6			16		
7			17		
8			18		
9			19		
10			20		

Table 5.6 Table for calculation of Mann-Whitney test

- Insert data into table, calculate ranks and insert into table (see 'Make-believe example of Mann-Whitney test for context of retrieval', page 36).
- Find the sum of ranks for the smaller sample (T). If both samples are of equal size, find the sum of the samples in condition A.
- Then, to find U, multiply N_A by $N_B = 10 \times 10 = 100$, where N_A is the number of participants in condition A and N_B is the number of participants in condition B.
- Add this number to $N_A \times (N_A + 1) \div 2 - T$
- Then find $U' = N_A \times N_B - U$
- The smaller of U and U' is compared to the *cv*.
- Find the *cv* from an appropriate critical value table – this will be dependent on whether you have a one- or two-tailed hypothesis (directional or non-directional), the level of significance used and the value of N (the number of ranked pairs).
- If the smaller of U and U' is equal to or less than the *cv*, it is significant, and the null hypothesis can be rejected.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	–	–	–	–	–	–	–	–	0	0	0	1	1	1	2	2	2	2	3	3
4	–	–	–	–	–	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	–	–	–	–	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	–	–	–	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	–	–	–	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	–	–	–	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	–	–	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	–	–	0	2	4	6	9	11	13	16	18	21	24	26	33	31	34	37	39	42
11	–	–	0	2	5	7	10	13	16	18	21	24	27	30	39	36	39	42	45	48
12	–	–	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	–	–	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	–	–	1	4	7	11	15	18	22	26	31	34	38	42	46	50	54	58	63	67
15	–	–	2	5	8	12	16	20	24	29	34	37	42	46	51	55	60	64	69	73
16	–	–	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	–	–	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	–	–	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	–	0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	–	0	3	8	13	18	24	30	36	42	58	54	60	67	73	79	86	92	99	105

Table 5.7 Mann-Whitney: Critical value table of *U* for a one-tailed (directional) test at $p = 0.005$ and two-tailed (non-directional) test at $p = 0.01$. Dashes indicate no decision is possible at the stated level of significance. For any *N1* and *N2*, the observed value of *U* will be significant if it is equal to or less than the critical values shown

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	1	1
3	–	–	–	–	–	–	0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	–	–	–	–	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	–	–	–	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6	–	–	–	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7	–	–	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8	–	–	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9	–	–	1	3	5	7	9	11	14	16	18	21	23	26	28	31	3	36	38	40
10	–	–	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11	–	–	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12	–	–	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13	–	0	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14	–	0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15	–	0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16	–	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17	–	0	4	8	13	18	23	28	33	38	43	49	55	60	66	71	77	82	88	93
18	–	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19	–	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20	–	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

Table 5.8 Mann-Whitney: Critical value table of *U* for a one-tailed (directional) test at $p = 0.01$ and two-tailed (non-directional) test at $p = 0.02$. Dashes indicate no decision is possible at the stated level of significance. For any *N1* and *N2*, the observed value of *U* will be significant if it is equal to or less than the critical values shown

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	1	1
3	–	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	–	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	–	0	1	3	5	6	7	8	10	11	13	14	16	17	19	21	22	25	27
7	–	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	–	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	–	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	–	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	–	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	–	1	4	7	11	14	18	22	26	29	33	37	41	45	49	55	57	61	65	69
13	–	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	74	76
14	–	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	–	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	76	80	85	90
16	–	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	–	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	–	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	–	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	–	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Table 5.9 Mann-Whitney: Critical value table of *U* for a one-tailed (directional) test at *p = 0.025* and two-tailed (non-directional) test at *p = 0.05*. Dashes indicate no decision is possible at the stated level of significance. For any *N1* and *N2*, the observed value of *U* will be significant if it is equal to or less than the critical values shown

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0
2	–	–	–	–	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3	–	–	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4	–	–	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5	–	0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6	–	0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7	–	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8	–	1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9	–	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10	–	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	54	58	62
11	–	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12	–	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13	–	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14	–	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15	–	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16	–	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17	–	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18	–	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	–	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	–	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

Table 5.10 Mann-Whitney: Critical value table of *U* for a one-tailed (directional) test at *p = 0.05* and two-tailed (non-directional) test at *p = 0.10*. Dashes indicate no decision is possible at the stated level of significance. For any *N1* and *N2*, the observed value of *U* will be significant if it is equal to or less than the critical values shown

The Wilcoxon signed-matched ranks test

Criteria for choice

- The Wilcoxon signed-matched ranks test is used when a repeated measures design has been used and the data is of at least ordinal level.

Rationale of the test

- The test ascertains whether there is a significant difference in paired observations (measurements taken from the same participant). Differences in pairs of scores are ranked according to their size, with the sum of the ranks of the less frequent sign (positive or negative differences) assessed to see whether it differs significantly from a critical value. For example, seeing whether there is a significant difference in the highest number of sequences recalled for letters compared to numbers when both have been assessed in participants (see the Jacobs (1887) mini-practical, page 23).

How to calculate the Wilcoxon signed-matched ranks test

- Construct an appropriate table for the calculation of a Wilcoxon signed-matched ranks test.

Participants' scores for the first condition of the IV	Participants' scores for the second condition of the IV	Differences between scores	Ranks of scores

Table 5.11 Table for the calculation of a Wilcoxon signed-matched ranks test

- Data is put into the table with differences between pairs of scores ranked (with no attention paid to whether differences are plus (+) or minus (–)) and scores of 'no difference' omitted (see table for Jacobs (1887) mini-practical, page 23).
- Calculate *T*, which is the sum of ranks for the less frequent sign (either minus or positive).
- Find the *cv* from an appropriate critical value table – this will be dependent on whether you have a one- or two-tailed hypothesis (directional or non-directional), the level of significance used and the value of *N* (the number of ranked pairs).
- Compare *T* to the *cv*.
- If *T* is equal to or lower than the *cv*, there is a significant difference and the null hypothesis can be rejected.

Level of significance for a two-tailed (directional) hypothesis				
	0.10	0.05	0.02	0.01
Level of significance for a one-tailed (non-directional) hypothesis				
N	0.05	0.025	0.01	0.005
5	0			
6	2	0		
7	3	2	0	
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68

Table 5.12 Critical values of *T* for the Wilcoxon signed-matched ranks test. Values of *T* that are equal to or less than the table value are significant

The independent *t*-test

Criteria for choice

The independent *t*-test is used when an independent groups design has been used and the data is of interval/ratio level.

Rationale of the test

The test compares the size of the differences in the mean scores of two sets of data drawn from independent (non-related) sources to see if they differ significantly from each other. For example, to see if the score on a test done by participants having no sleep last night differs significantly from those doing the test after eight hours' sleep last night (see page 12).

How to calculate the independent *t*-test

A make-believe example relating to scores on a test performed after conditions A and B of no sleep or eight hours' sleep last night will be used to demonstrate how to calculate the test.

- Construct an appropriate table for the calculation of an independent *t*-test.

Participant	Scores on test for condition A (no sleep)	A scores²	Participant	Scores on test for condition B (sleep)	B scores²
1	3	9	8	6	36
2	5	25	9	5	25
3	2	4	10	7	49
4	4	16	11	8	64
5	2	4	12	9	81
6	6	36	13	4	16
7	7	49	14	7	49
			15	8	64
			16	9	81
			17	7	49

Table 5.13 Table for the calculation of an independent *t*-test

1 A: add all A scores together (no sleep)
= 29

2 A: divide sum of A scores by number of participants in condition A (N_A)
= 29 ÷ 7
= 4.14

3 A: square each of the A scores (see Table 5.13)

4 A: add the squares of the A scores together
= 143

5 A: square the total of all A scores added together
= 29^2
= 841

6 A: divide the total of all A scores added together squared by the number of participants in condition A
= 841 ÷ 7
= 120.1

7 A: subtract 120.1 from the total of the squares of the A scores added together
= 143 – 120.1
= 22.9

8 Repeat steps 1–7 for the B scores (as steps 9–15)

9 B: 70

10 B: (N_B)
= 70 ÷ 10
= 7

11 B: (see Table 5.13)

12 B: 514

13 B:
= 70 × 70
= 4900

14 B:
= 4900 ÷ 10
= 490

15 B:
= 514 – 490
= 24

16 Add the scores from steps 7 and 15 together
= 22.9 + 24
= 46.9

17 Divide the result of step 16 by $N_A - 1$ added to $N_B - 1$
= 46.9 ÷ 6 + 9
= 46.9 ÷ 15
= 3.13

18 Find the reciprocal of N_A and N_B and add them together
$= \frac{1}{7} + \frac{1}{10}$
= 0.1429 + 0.1
= 0.2429

19 Multiply result of step 17 by result of step 18
= 3.13 × 0.2429
= 0.76

20 Find the square root of result of step 19
= √ 0.76
= 0.872

21 Take result of step 10 from result of step 2
= 4.14 – 7
= –2.86

22 Divide result of step 21 by result of step 20 to find *t*
= –2.86 ÷ 0.872
= –3.28

- Find the *cv* from an appropriate critical value table – this will be dependent on whether you have a one- or two-tailed hypothesis (in this case it is two-tailed/non-directional), the level of significance used (in this case 0.05) and the number of degrees of freedom (*d.f.*), which can be calculated from *d.f.* = $N_A + N_B - 2$ (which in this case = 7 + 10 – 2 = 15).
- If *t* is equal to or greater than the *cv* (which in this case is 2.131), there is a significant difference and the null hypothesis can be rejected. (Remember: whether a *t*-value is negative or positive is ignored.)

Level of significance for a one-tailed (directional) hypothesis			
d.f.	**0.1**	**0.05**	**0.025**
Level of significance for a two-tailed (non-directional) hypothesis			
d.f.	**0.2**	**0.1**	**0.05**
1	2.0	6.314	12.706
2	1.895	2.92	4.303
3	1.644	2.353	3.182
4	1.533	2.132	2.776
5	1.487	2.015	2.571
6	1.446	1.943	2.447
7	1.41	1.895	2.365
8	1.4	1.860	2.306
9	1.389	1.833	2.262
10	1.376	1.812	2.228
11	1.368	1.796	2.201
12	1.364	1.782	2.179
13	1.358	1.771	2.16
14	1.355	1.761	2.145
15	1.349	1.753	2.131
16	1.343	1.746	2.12
17	1.338	1.74	2.110
18	1.336	1.734	2.101
19	1.334	1.729	2.093
20	1.332	1.724	2.086
21	1.328	1.721	2.08
22	1.327	1.717	2.074
23	1.325	1.714	2.069
24	1.323	1.711	2.064
25	1.321	1.708	2.06
26	1.318	1.706	2.056
27	1.316	1.703	2.052
28	1.314	1.701	2.048
29	1.312	1.699	2.045
30	1.31	1.697	2.042

Table 5.14 Critical value table for the *t*-test (independent and related *t*-tests)

To be significant, *t* should be equal to or greater than the table value.

Degrees of freedom (*d.f.*) for a related *t*-test = $N - 1$.

Degrees of freedom for an independent *t*-test = $N_1 + N_2 - 2$.

The repeated t-test

Criteria for choice

The repeated *t*-test is used when a repeated design has been used and the data is of interval/ratio level.

Rationale of the test

The test compares the size of the differences in the mean scores of two sets of data drawn from related sources (not independent sources) to see if they differ significantly from each other. For example, seeing if the score on a test done by participants after meditating differs significantly from those doing the test without previously meditating.

How to calculate the repeated *t*-test

A make-believe example relating to scores on a test performed after meditating (condition A) and without previously meditating (condition B) will be used to demonstrate how to calculate the test.

- Construct an appropriate table for calculation of a repeated *t*-test.

Participant	Condition A: test scores after meditation	Condition B: test scores without meditation	*d*: sum of differences between pairs of scores	d^2: squares of sum of differences between pairs of scores
1	3	6	3	9
2	8	14	6	36
3	4	8	4	16
4	6	4	−2	4
5	9	16	7	49
6	2	7	5	25
7	12	19	7	49

Table 5.15 Table for the calculation of a repeated t-test

1. Calculate the difference between each pair of scores (B – A) and place into table
2. Add all the differences between pairs of scores together
   = 30
3. Divide the result from step 2 by the number of pairs of scores (*N*)
   = 30 ÷ 7
   = 4.29
4. Square the differences between pairs of scores and add all the squares together
   = 188
5. Square the result from step 2 and divide this number by *N* (the number of pairs of scores)
   = 30 × 30 ÷ 7
   = 900 ÷ 7
   = 128.57
6. Subtract the result of step 5 from the result of step 4
   = 188 – 128.57
   = 59.43
7. Divide the result from step 6 by *N*(*N* – 1)
   = 59.43 ÷ 42
   = 1.41
8. Find the square root of the result from step 7
   = $\sqrt{1.41}$
   = 1.19

9 Find t by dividing the result from step 3 by the result from step 8
$= 4.29 \div 1.19$
$= 3.61$

- Find the *cv* from an appropriate critical value table – this will be dependent on whether you have a one- or two-tailed hypothesis (in this case it is two-tailed/non-directional), the level of significance used (in this case 0.05) and the number of degrees of freedom (*d.f.*), which can be calculated from *d.f.* = $N - 1$, (which in this case = $7 - 1 = 6$).
- If t is equal to or greater than the *cv* (which in this case is 2.45), there is a significant difference and the null hypothesis can be rejected. (Remember: whether a t-value is negative or positive is ignored.)