CHAPTER 6

SIMPLE EXPERIMENTAL DESIGNS: BEING WATCHED

Contents

Simple Experimental Designs: Being Watched

An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts – for support rather than for illumination. (Andrew Lang)

Learning Objectives

- To examine the way in which data can be derived in simple experimental designs, and what those designs might consist of.
- To consider the elements of simple between- and within-subjects designs.
- To explain analysis of data derived from each of these two types of design.
- To explain the alternative analyses for parametric and non-parametric cases.

KEY TERMS

- Between subjects (independent measures)
- Control condition
- Directional hypothesis
- Experimental condition
- Mann–Whitney *U* test
- One–and two-tailed tests.
- Tied ranks
- *t*-test (unrelated/independent and related measures)
- Wilcoxon test
- Within subjects (related measures)

The simplest form of experimentation compares behaviour under two conditions that vary only by one thing; all aspects of the two situations are the same (or as similar as can be achieved), except for one single change. So, for example, people taking part in an experiment, the participants, may carry out some task, but in one condition they are alone (the control condition) and in the other they are being watched (the experimental condition). Any difference noted between the two conditions, all other elements being identical, can be attributed to the manipulation that has been performed. The manipulation here is on the independent variable of being watched, which has two conditions, alone or observed. What we are investigating is the effect of the changes in the independent variable on the dependent variable, the task performance.

WHO IS WATCHING YOU?

Have you ever noticed that you do things differently when you are being watched? You try a little harder, persevere a little longer, especially if it is something at which you are good. This enhancing effect of an audience is called *social facilitation*. Zajonc (1965) suggested a drive theory explanation for the effect of being watched. A 'drive' is a directed need for something, which must be reduced in order to satisfy some internal state. For example, hunger is a drive that can be reduced by eating, and anxiety is a drive that can be reduced by certain behaviours appropriate to the situation. KT Zajonc's **drive theory** states that, as we never know what an audience's response will be, we need to be in a high state of arousal when performing, or that the heightened arousal is simply an instinctive reaction to the presence of other people. This arousal results in a drive, which is reduced by the dominant response to the situation, utilising skills as best we can. The earliest scientific observation of social facilitation recorded was by Triplett in 1897. He observed that bike racers were more likely to cycle faster when this was against another cyclist than against the clock. This seems to be a nice, simple, clear-cut finding, but hardly explains the sometimes adverse effect of performance anxiety (stage fright). It does appear that social facilitation is not the straightforward explanation for audience effects it was first assumed to be. Later researchers also noted that an audience could have a detrimental effect on performance, particularly when the task is complex and/or unfamiliar. Other possible reasons for this social inhibition can be the perceived evaluation from the audience. Such evaluation apprehension (Cottrell, 1972) means that anxiety about the evaluation should increase arousal, which should increase performance. But Sanna and Shotland (1990) found that if a performer thinks that the audience is going to evaluate negatively, then this would have a damaging effect in comparison with a perceived positive evaluation. A further complication comes from Baron's (1986) attentional view of social facilitation. He suggested that an audience could have a distracting effect on the performer, with task performance being dependent upon the number of cues or distractions present. We have all seen athletes spurred on by the crowd shouting, but that if we add in inappropriate cues, such as a fan running onto the track, the athletes would not perform quite as well.

So, we appear to have a highly complex social effect that would not lend itself to simple experimentation. However, what Zajonc was suggesting is that it is the dominant response that is affected by the mere presence of an observer. In other words, the things that you already do well or automatically, such as riding a bike, throwing a ball, or dancing a tango, will be enhanced when you are watched. This is now the more up-to-date definition of social facilitation: that there will be strengthening of dominant, or well-learned, responses due to the presence of others. What is important here is not whether the effect of the audience's perceived evaluation will affect performance, but Zajonc's assertion that the mere presence of an audience will have the effect of arousing participants so much that they will at least attempt to perform to the best of their ability. This is the most basic effect of an audience that can be examined, and, if this effect is seen to not be so, the premise on which other theories of social facilitation is based falls down.

The simplest psychological experimental design compares results from two conditions, one in which some manipulation has happened and the other, the control condition, in which no such manipulation has taken place. If the two conditions have resulted in statistically significant different performance, then we can conclude that the manipulation of the independent variable has had an effect. In our social facilitation situation, we would have exactly the same setup for our participants; they will perform the task in the same environment except in one set of trials they will be watched. This is the classical design of an experiment and has certain components.

Firstly, there is direct comparison between two sets of circumstances or conditions; participants are alone or there is at least one other person present.

Secondly, there is control – the two conditions are as equal as we can make them except for the item we are investigating, the presence of another person or other people.

However, there are different variations even of this simple design. The description above of the effect of the audience is termed a repeated measures design as both conditions include data collected from the same person; in other words, any individual is measured both with and without observation. There are other terms for this, such as related measures or **within-subjects design**. The latter is the one we will use throughout the rest of the chapters. It describes measurements made *within* a group of people, whereas **between-subjects design** refers to measurements made *between* two groups of people. This second set of circumstances is also referred to as independent measures or unrelated samples.

KT

KT

So a simple experiment can have a within-subjects or between-subjects design. Usually simple designs refer to experiments in which there are only two conditions, so those with more than two conditions are addressed in a later chapter.

THE ANALYSIS OF DATA FROM EXPERIMENTS WITH TWO CONDITIONS

If the data derived from measures in the dependent variable conforms to parametric assumptions (see Chapter 5) then in order to decide if the means in the two conditions are significantly different a statistical test called a ***t* test** can be performed. The *t* test assesses whether the means of two sets of data are statistically different from each other. This analysis is appropriate whenever we want to compare the means of two distributions of data, and especially appropriate as the analysis for the two-sample experimental design.

What does it suggest if we say that the means for two conditions are statistically different? Well, if the experiment has been performed properly, we should be confident that it means that the two conditions differ because of something we have done to one to make it different from the other. However, consider Figure 6.1, in which the difference in the means is the same, but the distributions appear very different.

The top graph shows a case with moderate variability of scores within each group of data scores. The second situation shows the high-variability case and the third shows the case with low variability. We would conclude that the two groups appear most different in the low-variability case, as there appears to be relatively little overlap between the two curves. In the high-variability case, the group difference appears least striking because the two distributions overlap so much. So when we examine the differences between scores for two groups, we have to judge the difference between their means relative to the spread or variability of their scores. The *t* test does just this.

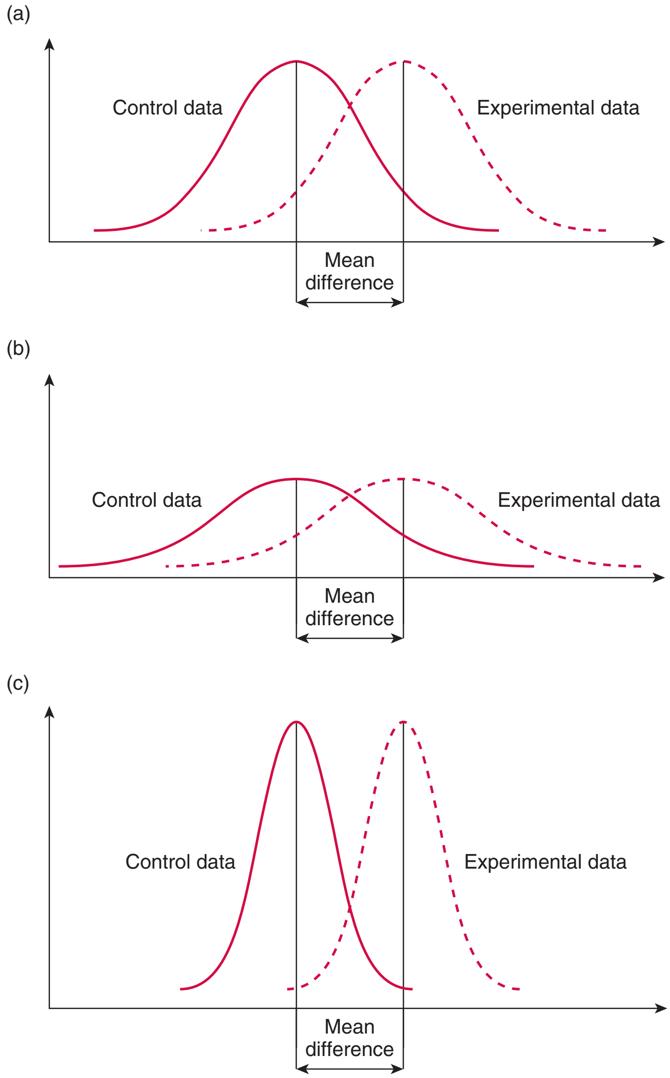

Figure 6.1

The *t* Test

The formula for the t test is a ratio:

$$t = \frac{\text{difference between group means}}{\text{variability of groups}}$$

The top part of the ratio is just the difference between the two means or averages, a measure of how different our treatment has made the two groups. The bottom part is a measure of the variability or dispersion of the scores. Essentially what this says is that the value of interest is the difference between the average scores in the two sets of data, but that we need to take into account any variability due to individual differences, the variance.

The *t* value will be positive if the first mean is larger than the second and negative if it is smaller. Once we have found our t-value we need to look it up in a table of significance to test whether the ratio is large enough to conclude that the difference between the groups is not likely to have been a chance finding. To test the significance, we need to set a level at which we will accept that the difference is significant, the significance level or alpha level (α). This level and its use are described in detail in Chapter 5, but conventionally we set this at 0.05 (or 0.01). We also need to determine the degrees of freedom (*df*) for the test. Given the alpha level, the *df* and the *t*-value, we can look the *t*-value up in a standard table of significance to determine whether the *t*-value is large enough to be significant. If it is, we can conclude that the difference between the means for the two groups is different. Fortunately, statistical computer programs routinely print the significance test results and save us the trouble of looking them up in a table.

EXPERIMENTS WITH WITHIN SUBJECTS DESIGNS

So, we have set up the social facilitation study to measure the dominant response of a group of people in the two conditions, one where they perform alone and one where they are watched by one other person. The dominant response, according to Platania and Moran (2001), is one where the participant is not judged in terms of correct or incorrect responses, but can give responses without the effects of competition, imitation, reinforcement, rewards or punishments. In their experiment the task was a stimulus discrimination task in which stimuli consisted of 11 squares ranging from 22 mm to 52 mm, increasing by increments of 3 mm. The participants were told that they were to distinguish between 11 different-sized squares, calling the smallest square in the trial '1' and the largest '11'.The remaining numbers (2 to 10) were to be assigned to the remaining squares in order of increasing size. In a practice session the participants were shown a guide sheet displaying the 11 squares in ascending order, and then were presented with slides of each square in serial order, from smallest to largest. They were told the correct number to be assigned to the square, and then told that the squares would appear in random order.

The recorded response was the number of times that each participant used his or her two preferred response numbers (responses with the greatest habit strength). A mean of 12 would be obtained if the participant were responding randomly or, alternately, accurately (i.e. two preferred responses, or numbers of particular squares multiplied by six trials). It does not matter if the participant is correct or not. The resulting data was two samples of discrimination responses, A and B.

Analysis of Data in Within Subjects Designs

Here we are going to use two sets of data from 24 participants, each participant taking part in both conditions, $N_{observed} = 24$ and $N_{alone} = 24$. We are going to treat this data as interval, and assume that the distribution of responses would be normal in the population, therefore we can use a parametric test.

Data Conforms to Parametric Assumptions

Table 6.1 shows the data from each participant in both conditions. The null hypothesis here is that the mean response is unrelated to whether the participant is observed, but, on average, participants produced more dominant responses when observed than when alone. What does this difference mean statistically?

We are concerned only with the *difference* between the observed and alone conditions, so we can treat this as one sample of data described as difference (D). Table 6.2 shows the same data we saw earlier, but now with the calculation of D for each subject. The mean of all these D-values is the same as the difference between the means above, but there is a smaller measure of variability. The square of D is also computed, as we will need that later.

If there were no tendency for people to perform differently under the two conditions, we would expect the mean of the D-values in such a sample to be zero. So is the observed mean difference of 3.16 significantly different from zero? We need only do a simple set of calculations:

$$t = \frac{\bar{x}_{\text{condition1}} - \bar{x}_{\text{condition2}}}{\sqrt{\frac{(\sum D^2) - \left((\sum D)^2/N\right)}{N(N-1)}}}$$

$$\begin{aligned} t &= (8.7083 - 5.4583) \div \sqrt{[(399.5 - (78^2/24)/(24 \times 23)]} \\ &= 3.25 \div \sqrt{146/552} = 3.25 \div 0.514288 \\ &= 6.319 \end{aligned}$$

Table 6.1

Participant	Observed	Alone
1	9.5	1
2	6.5	4
3	10.5	6.5
4	11.5	6.5
5	12	4.5
6	12	4
7	3	3
8	9.5	7
9	12	11
10	6.5	5
11	8	2.5
12	4.5	5
13	1	0
14	10.5	5
15	7.5	4
16	12	8
17	10	9
18	10.5	7.5
19	9	6.5
20	11	8
21	1	1
22	12	7
23	7	6
24	12	9
Mean	8.71	5.46
SD	3.4544	2.7462

Now we need to refer the calculated value of t to the table of critical values of t with $df = N-1$. In our example, $t = +6.319$, $df = 23$, which is well beyond the critical value for t.

Data Does Not Conform to Parametric Assumptions

Like the t test for related samples, the Wilcoxon signed-ranks test applies to two- sample designs involving within-subjects measures. However, this non-parametric test can be

Table 6.2

Participant	Observed	Alone	*D*	D^2
1	9.5	1	8.5	72.25
2	6.5	4	2.5	6.25
3	10.5	6.5	4	16
4	11.5	6.5	5	25
5	12	4.5	7.5	56.25
6	12	4	8	64
7	3	3	0	0
8	9.5	7	2.5	6.25
9	12	11	1	1
10	6.5	5	1.5	2.25
11	8	2.5	5.5	30.25
12	4.5	5	−0.5	0.25
13	1	0	1	1
14	10.5	5	5.5	30.25
15	7.5	4	3.5	12.25
16	12	8	4	16
17	10	9	1	1
18	10.5	7.5	3	9
19	9	6.5	2.5	6.25
20	11	8	3	9
21	1	1	0	0
22	12	7	5	25
23	7	6	1	1
24	12	9	3	9
Mean	8.7083	5.4583	3.25	16.6458
SD	3.4544	2.7462		
Sum			78	399.5

applied if we conclude that the data may not conform to parametric assumptions. Whilst the t test is robust enough to withstand violations of parametric assumptions, some researchers prefer to use non-parametric tests in smaller samples of data. Smaller samples mean we cannot always be sure that the data will fall into a normal distribution, or that there will be homogeneity of variance. In our sample above, we cannot be certain the distribution is normal even with Kolmogorov–Smirnov results that are not significant (see Chapter 5). Although we have treated the data as interval, we cannot be sure that it conforms to all parametric assumptions, therefore we might want to do a non-parametric test. The **Wilcoxon test** is a non-parametric test that compares two sets of data that are derived from the same people.

KT

The procedure is similar to the *t* test in that we are comparing differences, but once we have worked out the differences, we need to rank them in order, ignoring the sign (positive or minus), and disregarding differences of zero (see Table 6.3).

If there are numbers in *D* that are the same they are given a mean rank, so the ranks of 1, where there are three of them, would normally be given the ranks of 1, 2, 3 and 4, but are assigned 2.5. Similarly the differences of 1.5 would normally have ranks 5, 6 and 7, so are given (5 + 6 + 7)/3.

Once we have the number ranked we need to reassign the plus or minus sign from the original difference. Each of these two groups, the negative and positive ranks, are summed; the smaller one of these summed ranks is Wilcoxon's *T* . The output from a computer program calculation for Wilcoxon's *T* is shown in Table 6.4.

This value is compared with the critical values of *T*. Again we see that this is significant. Note that it is a minus number. This is simply due to the sample in which the highest ranks appeared, but could have implications in a parametric test. See the section on one-tailed and two-tailed tests below.

EXPERIMENTS USING BETWEEN-SUBJECTS DESIGNS

Between-subjects designs are used when the effect of the independent variable on the dependent variable might be such that a within-subjects design is inappropriate. For example, giving a test of memory to a group of people, then giving them alcohol and retesting them, might not show us the effect of the alcohol on memory because the people might have learnt the items on the memory test. So a between-subjects design would be used, and the memory test would be given to a group of people that had not received any alcohol and to one that had. Assuming that we can say that all other aspects are equal, such as age group, intellectual ability, etc., then any difference can be attributed to the effect of the alcohol. If we were to use the between-subjects design in our social facilitation experiment it would be because we thought the task could be learnt between the two sets of conditions. This is unlikely, and it would not be appropriate to use the between-subjects design. However, simply because we have a significant paired samples *t* test result, this does not mean that we can accept without question the effect of observation. There may be other effects

Table 6.3

Participant	Observed	Alone	*D*	Rank *D*
1	9.5	1	8.5	22
2	6.5	4	2.5	6
3	10.5	6.5	4	14.5
4	11.5	6.5	5	16.5
5	12	4.5	7.5	20
6	12	4	8	21
7	3	3	0	
8	9.5	7	2.5	9
9	12	11	1	2.5
10	6.5	5	1.5	6
11	8	2.5	5.5	18.5
12	4.5	5	–0.5	6
13	1	0	1	2.5
14	10.5	5	5.5	18.5
15	7.5	4	3.5	13
16	12	8	4	14.5
17	10	9	1	2.5
18	10.5	7.5	3	11
19	9	6.5	2.5	8
20	11	8	3	11
21	1	1	0	
22	12	7	5	16.5
23	7	6	1	2.5
24	12	9	3	11

on the social facilitation phenomenon that are dependent on the characteristic of the individual. For example, there may be a sex difference in the way that the performance changes between the observed and alone conditions. According to Steele's (1998) theory of stereotype threat, negative portrayals of individuals belonging to certain groups such as ethnic minorities or women can be internalised to such an extent that this state affects scores on ability tests and athletic performance. For example, the commonly held assumption that women are less spatially aware than

Table 6.4

		N	Mean Rank	Sum of Ranks
alone2 - observed2	Negative Ranks	21(a)	12.00	252.00
	Positive Ranks	1(b)	1.00	1.00
	Ties	2(c)		
	Total	24		

a alone2 < observed2
b alone2 > observed2
c alone2 = observed2

Test Statistics(b)

	alone2 – observed2
Z Asymp. Sig. (2-tailed)	–4.080(a) .000

a Based on positive ranks.
b Wilcoxon Signed Ranks Test

men can lead to lower performance on tasks that require such ability, like our comparative sizing task. It has been shown that when female participants are primed with negative stereotypes, scores on tests are significantly lower than if the women were led to believe the tests did not reflect these stereotypes (Spencer & Steele, 1999; Ben-Zeev, Fein & Inzlicht, 2005).

Fortunately, as seen in Table 6.5, we already were aware of this effect, and had recorded the sex of our participants along with everything else!

Analysis of Data in Between-Subjects Designs

Data Conforms to Parametric Assumptions

Here we need to look at any differences between male and female participants in the difference data. So our data now looks like that in Table 6.6.

There appears to be a bigger change between observed and alone conditions when the participants are male. Is this a statistically significant difference?

Table 6.5

Sex	Participant	Observed	Alone	*D*
Male	1	9.5	1	8.5
Male	2	6.5	4	1.5
Female	3	10.5	6.5	4
Male	4	11.5	6.5	5
Male	5	12	4.5	7.5
Male	6	12	4	8
Female	7	3	3	0
Male	8	9.5	7	2.5
Female	9	12	11	1
Female	10	6.5	5	1.5
Male	11	8	2.5	5.5
Female	12	4.5	5	–1.5
Female	13	1	0	1
Male	14	10.5	5	5.5
Male	15	7.5	4	3.5
Male	16	12	8	4
Female	17	10	9	1
Female	18	10.5	7.5	3
Female	19	9	6.5	2.4
Female	20	11	8	3
Female	21	1	1	0
Male	22	12	7	5
Female	23	7	6	1
Male	24	12	9	3

The formula for the *t* test is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\left(\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2\right)}{(n_1 + n_2) - 2} \times \left(\frac{1}{n_2} + \frac{1}{n_2}\right)}}$$

This can be broken down into several simple steps. The top line is the difference between the mean change between conditions for the male and female participants:

$$5.0417 - 1.4583 = 3.5834.$$

$$n_1 + n_2 - 2 = 22 \text{ and } \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.16667$$

For our data, these calculations become

$$\begin{aligned} t &= 3.5834 \div \sqrt{((47.73 + 21.23) \div 22) \times 0.16667)} \\ &= -3.5834 \div \sqrt{0.52243 = 3.5834 \div 0.72279} \\ &= 4.9577 \end{aligned}$$

Levene's test for equality of variances shows that the *F*-value of the variances of the two samples is 1.585, which is not significant, therefore there can be assumed to be homogeneity of variance. In this case, *t* is calculated to be 4.958 with 22 degrees of freedom, which is significant at the 0.05% significance level.

Data Does Not Conform to Parametric Assumptions

KT

We need a test as an alternative to the independent samples *t* test when the assumption of normality or equality of variance is not met. The usual test to use is the **Mann–Whitney *U* test**. This, in line with other non-parametric tests, uses the ranks of the data rather than their raw values to calculate the statistic. Since this test does not

Table 6.6

Female	Male
4	8.5
0	1.5
1	5
1.5	7.5
–1.5	8
1	2.5
1	5.5
3	5.5
2.4	3.5
3	4
0	5
1	3

make a distribution assumption, it is not as powerful as the *t* test, but we can test similar hypotheses that the data come from different populations. The Mann–Whitney produces a *U*-value that can be compared with a table of critical values for *U* based on the sample size of each group. If *U* exceeds the critical value for *U* at our significance level (0.05) it means that there is evidence to reject the null hypothesis.

The *U* test is easily calculated by hand, especially for small samples. There are two ways of doing this depending on the size of the sample. Having ranked each sample, we call the sample for which the ranks are smaller, sample 1.

Taking each observation in sample 1, we count the number of observations in sample 2 that are smaller than it (count a half for any that are equal to it). The total of these counts is *U*.

For larger samples, a formula can be used:

1 Arrange all the observations into a single ranked series, i.e., rank by disregarding which sample they are in.
2 Add up the ranks in sample 1
3 'U' is then given by: $U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$

where n_1 and n_2 are the two sample sizes, and R_1 is the sum of the ranks in sample 1.

For sample sizes greater than 8, a *z*-value can be used to approximate the significance level for the test. In this case, the calculated *z* is compared with the standard normal significance levels. The output from a computer program running the Mann–Whitney *U* test is shown in Table 6.7.

ONE- AND TWO-TAILED *t* TESTS

KT

Hypotheses can indicate the direction of effects. In order to test whether a **directional hypothesis** is supported, a one-tailed test would be carried out. A one- or two-tailed *t* test is determined by whether the total area of α is placed in one tail or divided equally between the two tails. The one-tailed *t* test is performed if the results are relevant only if they turn out in a particular direction. The two-tailed *t* test is performed if the results are relevant in either direction. The choice of a one- or two-tailed *t* test affects the hypothesis testing procedure in a number of different ways.

A two-tailed *t* test divides α in half, placing half in each tail. The null hypothesis in this case is a particular value, and there are two alternative hypotheses, one positive and one negative. The critical value of *t*, t_{crit}, is written with both a plus

Table 6.7

	Sex	N	Mean Rank	Sum of Ranks
D	male	12	17.58	211.00
	female	12	7.42	89.00
	Total	24		

	D
Mann–Whitney U	11.000
Z	–3.536
Asymp.Sig (2-tailed)	.000

and minus sign (±). For example, the critical value of t when there are 10 degrees of freedom ($df = 10$), and α is set to 0.05, is $t_{crit} = \pm 2.228$.

There are really two different one-tailed t tests, one for each tail. In a one-tailed t test, all the area associated with α is placed in either one tail or the other. Selection of the tail depends upon which direction t_{obs} would be if the results of the experiment came out as expected. The selection of the tail must be made before the experiment is conducted and analysed.

If the t-value were positive and over the critical value, then significance would be found in the two-tailed and the *positive* one-tailed t tests. The one-tailed t test in the negative direction would not be significant, because α was placed in the wrong tail. This is the danger of a one-tailed t test.

If t is negative and above the critical value then significance would only be found in the *negative* one-tailed t test. If the correct direction is selected, it can be seen that one is more likely to reject the null hypothesis. The significance test is said to have greater power in this case.

The selection of a one- or two-tailed t test must be made before the experiment is performed. It is not acceptable practice to find that the t-value is over the critical value and then decide to do a one-tailed test. Readers of published articles are sometimes suspicious when a one-tailed t test is done; the recommendation is that if there is any doubt, a two-tailed test should be done.

Summary

We have looked at the analysis of data derived from simple experimental designs, in which there are two samples of data to consider. Data from simple experiments with a between-subjects design and a within-subjects design were considered. When a between-subjects design is employed and the data conforms to parametric conditions, the independent samples *t* test is appropriate and we have looked at how this is done, together with the alternative analysis when the data does not conform to parametric assumptions, the Mann–Whitney *U* test. We have also looked at what happens when the experiment uses a within-subjects design and examined the appropriate analysis for data conforming to the parametric assumption (the related samples *t* test) and the non-parametric equivalent (the Wilcoxon sign test).

Professor Robert Zajonc

Professor Robert Zajonc studied at the University of Michigan, obtaining his PhD in 1955, and gaining a professorship. He held the positions of Director of the Institute for Social Research and Director of the Research Center for Group Dynamics In 1994 he joined the faculty at Stanford University, where he is currently Professor Emeritus of Psychology. Professor Zajonc's has received a number of honours in recognition of his work in social psychology.

Professor Zajonc's contribution to psychology has been the study of basic processes of social behaviour, and his attempts to show that the relationship between emotion and cognition is a strong one. His research demonstrates that affective influences could take place in the absence of cognitive contributions.