CHAPTER 23

t-Test Variants: Crossover Tests, Equivalence Tests

Contents

CROSSOVER TRIALS

The main requirement for a paired t-test is that if both members of the pair are given the same treatment, the results would on average be the same. One form of paired trial, often used, is to give some subjects treatment A, and then after a waiting period give the same subjects treatment B. The null hypothesis is that the two treatments have the same effect, and if we reject the null hypothesis there is some basis for concluding that one treatment is better than the other. An example might be giving a group of hypertensive subjects drug A for a few days, recording the change in blood pressure, and then a week later give the same subjects drug B for a few days to determine which drug caused the greater fall in pressure. For each subject there will be a pressure difference Δi. This difference would not be identical in each subject, and for each subject $\Delta i = \overline{\Delta} + \varepsilon_i$, where ε_i is the individual error term. These error terms have a mean of zero, and their variability allows the calculation of the standard deviation and standard error.

The concern with this design is that the effects of the first treatment might still be present when the second treatment is given. There might be residual blood levels of the drug, some receptors might still be occupied, psychological effects might alter responses, or some long-term physiological changes might have been caused. Some diseases get better or worse with time. If there is any such carryover effect, then any difference between treatments A and B is a function of a possible real effect of the drugs plus an unknown effect of time, and these cannot be separated. To solve some of these problems, the crossover design can be used, as described by Hills and Armitage, 1979.

Patients are randomized into two similar groups A and B: group A is given treatment X and group B is given treatment Y. One of the treatments can be a placebo. After an

Biostatistics for Medical and Biomedical Practitioners
http://dx.doi.org/10.1016/B978-0-12-802387-7.00023-8

Table 23.1 Basic 2 × 2 crossover trial

Period	Group A	Group B
1	Treatment X	Treatment Y
	Result	Result
	Washout	Washout
2	Treatment Y	Treatment X
	Result	Result

appropriate time to allow for washout of the effects of the treatment, the groups are reversed, so that group A gets treatment Y, and group B gets treatment X (Table 23.1).

The trial is designed to test two hypotheses: the mean values of treatments X and Y are significantly different, and there is no effect of time on the results. Among the assumptions required are similarity between patient groups, attained by randomization, and a response that is on average the same for the two periods on the same treatment; that is, the results of the treatments should not be affected by the order in which they were given (Brown, 1980; Hills and Armitage, 1979; Jones, 2008, 2010; Jones and Haughie, 2008; Jones and Kenward, 2003). For any given subject in group A, the response in period 1 is Y_1, and can be considered to be the sum of the fixed effect of the treatment T_X and a response that is due to the passage of time ε_{1A}. For that same subject, the response Y_2 in period 2 is $T_Y + \varepsilon_{2A}$. Similarly, a subject in the B group has responses in periods 1 and 2, respectively, $T_Y + \varepsilon_{1B}$ and $T_X + \varepsilon_{2B}$ (Table 23.2).

Table 23.2 Individual responses

Period	Group A subject	Group B subject
1	$Y_1 = T_X + \varepsilon_{1A}$	$Y_1 = T_Y + \varepsilon_{1B}$
2	$Y_2 = T_Y + \varepsilon_{2A}$	$Y_2 = T_X + \varepsilon_{2B}$

Based on the assumptions, the values of T_X and T_Y are fixed, but all the other values can change from subject to subject. The effect of treatment (the difference between X and Y) in a group A subject is determined from $d_A = Y_1 - Y_2 = (T_X - T_Y) + (\varepsilon_{1A} - \varepsilon_{2A})$, and in a group B subject it is $d_B = Y_2 - Y_1 = (T_X - T_Y) - (\varepsilon_{1B} - \varepsilon_{2B})$. If there is no effect of time, then the average values for $\varepsilon_{1A} - \varepsilon_{2A}$ and $\varepsilon_{1B} - \varepsilon_{2B}$ are zero; if there is an effect of time, then $\varepsilon_{1A} - \varepsilon_{2A} =$ some mean value δ, with standard error of $\frac{\sigma}{\sqrt{N_A}}$ for group A and $\frac{\sigma}{\sqrt{N_B}}$ for group B. Calculate the mean values for each group as $\overline{d_A}$ and $\overline{d_E}$, and then the average of the difference between these means is

$$\frac{\overline{d_A} - \overline{d_B}}{2} = \frac{\left\{\overline{(T_X - T_Y)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_A}\right\} - \left\{\overline{(T_X - T_Y)_B} - \overline{(\varepsilon_1 - \varepsilon_2)_B}\right\}}{2}$$

$$= \frac{\left\{\overline{(\varepsilon_1 - \varepsilon_2)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_B}\right\}}{2}$$

because the sums of $T_X - T_Y$ for each group cancel out. The standard error of this difference, as in the unpaired t-test, is $\frac{1}{2}\sqrt{\left(\frac{s_p^2}{N_A} + \frac{s_p^2}{N_B}\right)}$, and this can be used to determine whether it is possible to reject the hypothesis that $\overline{d_A} - \overline{d_B} = 0$, that is, that there is no average effect of time. If the null hypothesis is not rejected, then test the average effects of the two treatments as

$$\frac{\overline{d_A} + \overline{d_B}}{2} = \frac{\left\{\overline{(T_X - T_Y)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_A}\right\} + \left\{\overline{(T_X - T_Y)_B} - \overline{(\varepsilon_1 - \varepsilon_2)_B}\right\}}{2}$$
$$= \frac{\overline{(T_X - T_Y)_A} + \overline{(T_X - T_Y)_B}}{2}$$

because time has been shown to have no effects. This difference is tested for significant difference from zero by the same standard error. An alternative set of calculations and a simple explanation are provided by Wellek and Blettner (2012).

If there is a significant effect of time, then it might not be useful to proceed with the analysis. Various alternatives have been proposed, but care is needed in applying them (Jones et al., 1996). Some designs include three or more periods, for example, group A is given three successive treatments X, Y, Y and group B is given treatments Y, X, X (Ebbutt, 1984; Jones and Haughie, 2008; Laska et al., 1983). If the second and third identical treatments in each group are similar, it is unlikely that there is a carryover effect from the first treatment. As an example, Ramsey et al. (1993) studied the effect of aerosolized tobramycin in treating patients with cystic fibrosis who had pneumocystis infection. Group I was given aerosolized tobramycin for 28 days, followed by aerosolized half-normal saline for two 28-day periods. Group II was given aerosolized half-normal saline, followed by two periods of aerosolized tobramycin. The primary outcomes were based on tests of forced vital capacity, forced expiratory volume (FEV), and forced expiratory flow (FEF). Approximate differences from control values of FEF are shown in Table 23.3.

Table 23.3 Three period crossover trial

	Period 1	Period 2	Period 3
Group I	Placebo	Tobramycin	Tobramycin
	−7	+5	+4
Group II	Tobramycin	Placebo	Placebo
	+8	+1	+2

As shown, the duplicate values in periods 2 and 3 are almost identical, suggesting no carryover from period 1 to period 2. In this study, however, there was carryover for FEV.

Crossover designs can have more groups and can deal with ordinal numbers or binary categories (Brown, 1980). What are the advantages of the crossover design? Using each subject as his or her own control minimizes variability as compared with a parallel design with two groups, just as a paired t-test has less variability than an unpaired test because it does not have to allow for differences among subjects. Therefore the total number of subjects is less, often considerably less, for the crossover design. This is particularly important when studying treatments for a rare disease. Furthermore, unlike the paired test at two different times, the effect of time can be estimated.

Some key assumptions must be met for the crossover design to be useful (Brown, 1980; Jones, 2008; Hills and Armitage, 1979). (1) The two groups must be equally matched at the onset; it would be futile to have thin, nonhypertensive subjects in one group and obese hypertensive subjects in the other. (2) The subjects should be in the same clinical state at the beginning of the second period as they were at the beginning of the first period; that is, the first treatment should not leave the subject in a different state, and the disease process has not changed. (3) The effect of the agent used in the first treatment should not carry over to the beginning of the second period; that is, the drug or treatment activity should have a short half-life. (4) The order in which the treatments are given should not affect the results. Therefore crossover designs are best used for chronic diseases such as chronic obstructive pulmonary disease or rheumatoid arthritis. The design is not restricted to these chronic diseases, though. It has been used to test the ability of acetazolamide to prevent or modify mountain sickness, all of the above criteria being met (Greene et al., 1981). It has even been used to study the effect of sumatriptan on acute cluster headaches (Ferrari, 1991). Crossover designs are often used in equivalence studies.

As an example, treatment with acetazolamide in preventing acute mountain sickness was studied (Greene et al., 1981). Twenty-four amateur mountain climbers were divided at random into two groups. Before climbing Mt Kilimanjaro (5895 m) one group was given acetazolamide and the other a placebo. After descending, there was a 5-day rest period, and then the treatments were switched when the climbers ascended Mt Kenya (5186 m). Each climber made daily notes of symptoms, and a scoring system was used; the more symptoms, the higher the score. The results are given in Tables 23.4a and b.

The average effect due to time is

$$\frac{(4.83 - 1.91) - (2.5 - 14.25)}{2} = 7.34.$$

The average effect of the drug (difference between scores with acetazolamide and placebo) is

$$\frac{(4.83 - 1.91) + (2.5 - 14.25)}{2} = -4.42.$$

Table 23.4a Scores

	Group 1			Group 2		
	Acetazolamide Kilimanjaro (Period 1)	Placebo Mt Kenya (Period 2)	Period 1−2	Placebo Kilimanjaro (Period 1)	Acetazolamide Mt Kenya (Period 2)	Period 2−1
	7	0	7	25	−1	−26
	13	7	6	19	5	−14
	3	3	0	17	9	−8
	4	—	—	7	1	−6
	5	−1	6	9	3	−6
	6	−1	7	12	2	−10
	0	0	0	18	2	−16
	1	0	1	12	0	−12
	3	0	3	5	4	−1
	5	2	3	12	−1	−13
	9	9	0	18	−2	−20
	2	2	0	17	8	−9
ΣX	58	21	33	171	30	−141
$\overline{X}$	4.83	1.91	3	14.25	2.5	−11.75
N	12	11	11	12	12	12
s	3.61	3.30	3	5.74	3.50	6.7
$s_{\overline{X}}$			0.9			1.94

Data adapted from Greene et al. (1981).

Table 23.4b Summary of high-altitude trial results (see text)

Group	Treatment	Period	Mean score
I	A. Acetazolamide	1	4.83
	B. Placebo	2	1.91
II	A. Acetazolamide	2	2.5
	B. Placebo	1	14.25

From the data, the standard error was

$$\frac{1}{2}\sqrt{(0.9^2 + 1.94^2)} = 1.07.$$

Therefore to test the null hypothesis that time had no effect calculate $t = \frac{7.34}{1.07} = 6.86$. $P < 0.00001$, and we can safely reject the null hypothesis. This conclusion is reasonable because of the known effect of acclimatization to altitude. The effect of treatment can be tested by $t = \frac{4.42}{1.07} = 4.13$. $P < 0.00001$, also a reason to reject the null hypothesis. Many reports of crossover studies in the literature have ignored requirements (Baer et al., 2010).

N of 1 trials

These are variations of the cross-over trials in which treatments are given in random order to a single patient. All the issues about carry over pertain. If, for example, two medications for back pain are given and symptoms are recorded accurately, it might be possible to show that one treatment is better than the other *for that particular patient*. This avoids the "one size fits all" approach of randomized clinical trials with a gain in efficiency. (Lillie et al., 2011).

EQUIVALENCE AND NONINFERIORITY TESTING

One type of test that compares two samples seems to be the antithesis of a statistical test, and that is the equivalence or noninferiority test. Equivalence implies that the new mean is only slightly better or worse than the old mean, whereas noninferiority means that the new mean is not significantly worse than the old mean. These tests are aimed at introducing a new treatment that is cheaper, less invasive, has fewer side effects, or has other advantages (Pocock, 2003). A pharmaceutical company might want to establish the merits of a new preparation of a vaccine, or a new combination of vaccines. The advantages of the new preparation might be that it can be stored for longer times, or may save costs. What is important for the company is to show that the new preparation is not less effective than the old vaccine. If the new item is more effective than the old one that would be advantageous, but all that is required for the company to be licensed to produce the new vaccine is to show equivalent effectiveness with the previous one. This type of testing was found in 2% of vaccine studies reported in several major medical journals (Jacobson and Poland, 2005), but the principles apply widely. Other examples are examining the incidence of infective endocarditis before and after changing the guidelines for antibiotic prophylaxis (no difference was found) (Thornhill et al., 2011), comparing two different types of coronary stents (Hofma et al., 2012) or two types of stem cell transplantation (da Silva et al., 2008). The two latter references describe the methods of testing clearly.

Performing a standard t-test and finding that it does not disprove the null hypothesis is not a substitute for equivalence testing because it may merely reflect a low power. "Absence of evidence is not evidence of absence" (Altman and Bland, 1995). On the other hand, even a trivial difference between two means can be significant if the sample size is huge. What is important to consider is not significance but the effect size Δ. In noninferiority tests the investigator decides on what sized Δ is acceptable. For example, if drug A lowers blood pressure by a mean of 30 mm Hg, and drug B lowers blood pressure by a mean of 27 mm Hg, then drug B would be regarded as satisfactory. Many regulatory agencies accept a difference of as much as 15% of the mean, that is, if the new treatment is not more than 15% worse than the old treatment, then noninferiority (accepting the null hypothesis) may be asserted. One Federal standard accepts a 20% difference (Food and Drug Administration, 1977). It would be preferable to have a smaller deviation, for example, 5%, but this may demand an impractically large number of subjects.

Often the two one-sided test (TOST) is done to test the joint null hypothesis.

$$H_{01} : \mu_1 - \mu_2 \geq \Delta$$
$$H_{02} : \mu_1 - \mu_2 \leq -\Delta$$

Rejection of H_{01} implies that $\mu_1 - \mu_2 \leq \Delta$, and rejecting H_{02} implies that $\mu_1 - \mu_2 \geq -\Delta$. Rejecting both hypotheses implies that the difference lies within the range Δ to $-\Delta$ and hence that for practical purposes the two drugs have equivalent effects. Therefore do TOSTs.

$$t_1 = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \Delta}{S_{\overline{X_1} - \overline{X_2}}} \text{ and } t_2 = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (-\Delta)}{S_{\overline{X_1} - \overline{X_2}}}.$$

If neither t-test shows significance, then the observed difference lies within the permissible difference so that the two drugs have equivalent effects.

A variant of this test is to calculate confidence limits for the difference between the two means. If this lies within the limits $\pm\Delta$, which demarcates a zone of scientific or clinical indifference, equivalence is demonstrated. Figure 23.1, based on a similar figure by Jones et al. (1996), shows the principle.

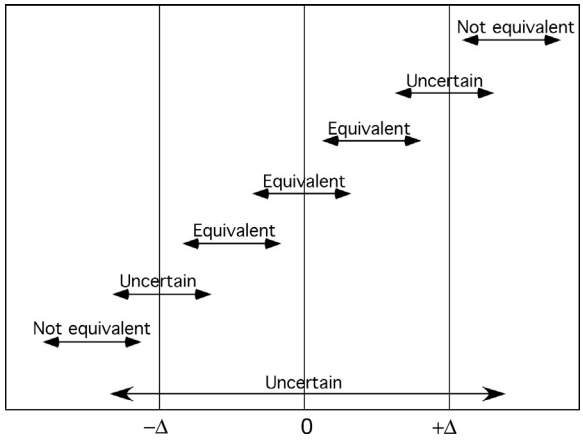

Figure 23.1 Range from $-\Delta$ to $+\Delta$ within which equivalence is assumed (zone of indifference).

Lines labeled "Equivalent" are within this range, so that studies producing these limits are equivalent to existing products. Lines labeled "Not-equivalent" are outside this range so that the two groups are not equivalent. Lines labeled "Uncertain" are inconclusive, and may call for further studies. Two of the lines showing confidence limits that demonstrate equivalence do not cross zero, so that they argue for rejecting the null hypothesis, but equivalence is still postulated because the observed difference is not meaningful. The European Agency for the Evaluation of Medicinal Products

has set out criteria for making these decisions at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003519.pdf and at http://www.tga.gov.au/pdf/euguide/ewp048299en.pdf. The confidence limits are calculated from $(\overline{X_1} - \overline{X_2}) \pm t_{0.10} S_{\overline{X_1} - \overline{X_2}}$, and $t_{0.10}$ is chosen so that the chances of rejecting the null hypothesis are 0.05 at each end of the limits. Tryon (2001) developed a method of inferential confidence intervals that slightly reduces the lengths of the confidence limits by about 10–20%.

One problem with equivalence testing is that it often requires large numbers of subjects because small differences are being examined. One estimate of numbers required is based on the formula

$$N \geq \frac{\left(z_\beta + z_\alpha\right)^2 s^2}{\Delta^2}.$$

Because twice z is squared, four times as many subjects are required as for a simple t- or z-test. Jacobson and Poland (2005) recommended a modified approach, based on a suggestion by Feinstein. This eliminates consideration of very small differences by setting two thresholds, one for an insignificant difference, designated i, and the other for the threshold of an important difference, Δ. Then it is possible to calculate

$$N \geq \frac{{z_\alpha}^2 s^2}{(\Delta - i)^2}.$$

What this does is to eliminate the need to consider trivial differences less than i (Deeks et al., 2005), with consequent reduction in the required numbers. A calculator for sample size can be found online at http://www.sealedenvelope.com/power/binary-noninferior.

A simple explanation of this subject can be found at http://www.graphpad.com/support/faqid/1061/.

REFERENCES

Altman, D.G., Bland, J.M., 1995. Absence of evidence is not evidence of absence. BMJ 311, 485.

Baer, H.J., Tworoger, S.S., Hankinson, S.E., Willett, W.C., 2010. Body fatness at young ages and risk of breast cancer throughout life. Am. J. Epidemiol. 171, 1183–1194.

Brown Jr., B.W., 1980. The crossover experiment for clinical trials. Biometrics 36, 69–79.

Deeks, J.J., Macaskill, P., Irwig, L., 2005. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J. Clin. Epidemiol. 58, 882–893.

Ebbutt, A.F., 1984. Three-period crossover designs for two treatments. Biometrics 40, 219–244.

Ferrari, M.D., 1991. Treatment of migraine attacks with sumatriptan. The subcutaneous sumatriptan international study group. N. Engl. J. Med. 325, 316–321.

Food and Drug Administration, 1977. The bioavailability protocol guideline for ANDA and NDA submission. Division of Biopharmaceutics, D. M. B. O. D., Food and Drug Administration.

Greene, M.K., Kerr, A.M., Mcintosh, I.B., Prescott, R.J., 1981. Acetazolamide in prevention of acute mountain sickness: a double-blind controlled cross-over study. Br. Med. J. (Clin. Res. Ed.) 283, 811–813.

Hills, M., Armitage, P., 1979. The two-period cross-over clinical trial. Br. J. Clin. Pharmacol. 8, 7–20.
Hofma, S.H., Brouwer, J., Velders, M.A., van't Hof, A.W., Smits, P.C., Queré, M., de Vries, C.J., van Boven, A.J., 2012. Second-generation everolimus-eluting stents versus first-generation sirolimus-eluting stents in acute myocardial infarction. 1-year results of the randomized XAMI (XienceV Stent vs. Cypher Stent in Primary PCI for Acute Myocardial Infarction) trial. J. Am. Coll. Cardiol. 60, 381–387.
Jacobson, R.M., Poland, G.A., 2005. Studies of equivalence in clinical vaccine research. Vaccine 23, 2315–2317.
Jones, B., 2008. The cross-over trial: a subtle knife. Significance 5, 135–137.
Jones, B., 2010. The waiting game: how long is long enough? Significance 2, 40–41.
Jones, B., Haughie, S., 2008. Cross-over trials in practice: tales of the unexpected. Significance 5, 183–184.
Jones, B., Jarvis, P., Lewis, J.A., Ebbutt, A.F., 1996. Trials to assess equivalence: the importance of rigorous methods. BMJ 313, 36–39.
Jones, B., Kenward, M.G., 2003. Design and Analysis of Cross-over Trials. Chapman & Hall/CRC, Boca Raton, FL.
Laska, E., Meisner, M., Kushner, H.B., 1983. Optimal crossover designs in the presence of carryover effects. Biometrics 39, 1087–1091.
Lillie, E.O., Patay, B., Diamant, J., Issell, B., Topol, E.J., Schork, N.J., 2011. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per. Med. 8, 161–173.
Pocock, S.J., 2003. The pros and cons of noninferiority trials. Fundam. Clin. Pharmacol. 17, 483–490.
Ramsey, B.W., Dorkin, H.L., Eisenberg, J.D., Gibson, R.L., Harwood, I.R., Kravitz, R.M., Schidlow, D.V., Wilmott, R.W., Astley, S.J., McBurnie, M.A., et al., 1993. Efficacy of aerosolized tobramycin in patients with cystic fibrosis. N. Engl. J. Med. 328, 1740–1746.
da Silva, G.T., Logan, B.R., Klein, J.P., 2008. Methods for equivalence and noninferiority testing. Biol. Blood Marrow Transpl. 15.
Thornhill, M.H., Dayer, M.J., Forde, J.M., Corey, G.R., Chu, V.H., Couper, D.J., Lockhart, P.B., 2011. Impact of the NICE guideline recommending cessation of antibiotic prophylaxis for prevention of infective endocarditis: before and after study. BMJ 342, d2392.
Tryon, W.W., 2001. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. Psychol. Methods 6, 371–386.
Wellek, S., Blettner, M., 2012. On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch. Arztebl. Int. 109, 276–281.