BMJ

BMJ 2014;348:g124 doi: 10.1136/bmj.g124 (Published 10 January 2014)

ENDGAMES

STATISTICAL QUESTION

Parametric statistical tests for two related groups: numerical data

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, London, UK

Researchers investigated whether *Ginkgo biloba* was effective in treating tinnitus.[1] A double blind placebo controlled trial study design was used. The intervention was 12 weeks' treatment with 50 mg *Ginkgo biloba* extract LI 1370 three times daily or placebo. Participants were 978 healthy people (489 matched pairs) aged 18-70 years with comparatively stable tinnitus. Pairs were matched for sex, age (within 10 years), and duration of tinnitus (within five years). Within each matched pair, one participant was allocated at random to the intervention and the other to placebo.

The main outcome measures included assessment of how loud and troublesome tinnitus was before and after treatment. The loudness and troublesome nature of tinnitus were rated on a six point and a five point scale, respectively. Assessment was made by mail and telephone. For each participant the change from baseline to end of treatment (12 weeks minus baseline) in each outcome measure was recorded. Distributional assumptions of normality in the outcome measures were verified. Paired data were compared between treatment groups.

There was no significant difference between the treatment groups (intervention minus placebo) in change over 12 weeks from baseline in loudness (mean −0.05 (standard deviation 1.48), 95% confidence interval −2.00 to 0.11; P >0.05) or troublesome nature (−0.10 (1.32), −0.24 to 0.04; P >0.05) of tinnitus. The researchers concluded that 50 mg *Ginkgo biloba* extract LI 1370 given three times daily for 12 weeks was no more effective than placebo in treating tinnitus.

Which one of the following statistical tests would most likely have been used to compare treatment groups in the change in the primary outcome measure of loudness over 12 weeks from baseline?

a) Paired *t* test

b) Student's *t* test

c) Wilcoxon rank sum test

d) Wilcoxon signed ranks test

Answers

The paired *t* test (answer *a*) would most likely have been used to compare the treatment groups in the change in the primary outcome measure of loudness over 12 weeks from baseline.

The aim of the trial was to establish the effectiveness of *Ginkgo biloba* in treating tinnitus. A randomised placebo controlled study design was used. Simple random allocation is typically used in trials to allocate participants to treatment after recruitment. Simple random allocation (commonly known as random allocation or randomisation) involves each participant having an equal probability of being allocated to each treatment group. If the sample size is large enough, random allocation will control for confounding at baseline by achieving treatment groups similar in those characteristics that influence the association between treatment and outcome. Confounding has been described in a previous question.[2]

In the trial above the participants were paired after recruitment. This involved finding two participants of the same sex, age (within 10 years), and duration of tinnitus (within five years). Within each matched pair, one participant was allocated to the intervention and the other to placebo, with allocation decided at random. Randomising participants within matched pairs to treatment groups is referred to as restricted randomisation, which has been described in a previous question.[3] Restricted randomisation is a method that controls the random allocation procedure to achieve greater equivalence between treatment groups in group size and baseline characteristics. The purpose of matching pairs of participants on sex, age, and duration of tinnitus was to reduce systematic differences between the treatment groups and therefore minimise confounding. Therefore, any differences between treatment groups in outcome would not result from differences in the matching variables. Sex, age, and duration of tinnitus were thought to have a substantial affect on the outcome measures.

The primary outcome measure was measured at baseline and 12 weeks later, with the change from baseline recorded for each participant. Within each matched pair, the difference between

p.sedgwick@sgul.ac.uk

treatment groups (intervention minus placebo) in the change from baseline was obtained. The paired *t* test (answer *a*) would most likely have been used to compare treatment groups in the mean change in the outcome measure of loudness over the 12 weeks of follow-up. The test is used to compare the means of two measurements made on the same or matched pairs of participants. The measurements for the matched pair are statistically dependent. Therefore, the matched pair of participants was the unit of analysis—not the single participant. The comparison of effects between treatments was made within each pair. The paired *t* test is a parametric test, and it assumed that the distribution of the differences between matched pairs in the change from baseline for the primary outcome in the population was normally distributed. The researchers confirmed that distributional assumptions of normality in the outcome measures were verified. Parametric tests have been described in a previous question.[4]

A 95% confidence interval for the mean difference between treatment groups in the change in the outcome measure of loudness over the 12 weeks of follow-up was presented. This is a further indication that the parametric paired *t* test (answer *a*) would most likely have been used to compare treatment groups. A 95% confidence interval for the mean difference should have been derived only if the assumption of normality required for the parametric paired *t* test could be made. If the assumption of normality could not have been made, the Wilcoxon signed ranks test (answer *d*), described further below, would have been used. Under such circumstances, it would not have been sensible to derive the 95% confidence interval for the mean difference between the treatment groups.

The paired *t* test used traditional hypothesis testing with a two sided alternative hypothesis, described in a previous question,[5] to compare treatment groups in the change from baseline in loudness of tinnitus. The null hypothesis states that in the population the mean difference between treatment groups in the mean change from baseline for the outcome of loudness of tinnitus was equal to zero. The alternative hypothesis states that the mean difference is not equal to zero.

The student's *t* test (answer *b*), also known as the independent samples *t* test,[6] compares the means of an outcome variable in two independent groups. It is a parametric test that assumes that the outcome variable is normally distributed in both groups and that the variances for the two groups are equal. Because the participants in the above trial were matched pairs the treatment groups were not independent. The student's *t* test would therefore not be the most appropriate test to compare the effects of treatment; in particular, it would not account for the paired nature of the data and the minimisation of confounding achieved through matching.

The Wilcoxon rank sum test (answer *c*) and Wilcoxon signed ranks test (answer *d*) are non-parametric methods that have been described in previous questions.[7 8] The Wilcoxon rank sum test is used to compare two independent groups in a variable measured on a continuous or ordinal scale. The Wilcoxon signed ranks test is used to compare two related samples in a variable that is continuous or ordinal; the two related samples may be two measurements made on the same or matched pairs of participants. The Wilcoxon rank sum test and Wilcoxon signed ranks test are non-parametric methods and therefore make no assumption about the distribution of the variable in the population. Non-parametric tests have been described in a previous question.[4] The tests are used when the distribution of the variable does not satisfy the assumption of normality or when it is not achieved after a transformation of the data. The log transformation of data has been described in a previous question.[9] Distributional assumptions of normality in the outcome measures were verified. Therefore, the paired *t* test would have been used in preference to the Wilcoxon signed ranks test.

Competing interests: None declared.

1 Drew S, Davies E. Effectiveness of Ginkgo biloba in treating tinnitus: double blind, placebo controlled trial. *BMJ* 2001;322:78.
2 Sedgwick P. Confounding in clinical trials. *BMJ* 2012;345:e7951.
3 Sedgwick P. Restricted randomisation. *BMJ* 2012;344:e1324.
4 Sedgwick P. Parametric v non-parametric statistical tests. *BMJ* 2012;344:e1753.
5 Sedgwick P. Statistical hypothesis testing. *BMJ* 2010;340:c2059.
6 Sedgwick P. Independent samples t test. *BMJ* 2010;340:c2673.
7 Sedgwick P. Non-parametric statistical tests for independent groups: numerical data. *BMJ* 2012;344:e3354.
8 Sedgwick P. Non-parametric statistical tests for two related groups: numerical data. *BMJ* 2012;344:e2537.
9 Sedgwick P. Log transformation of data. *BMJ* 2012;345:e6727.

Cite this as: *BMJ* 2014;348:g124