BMJ 2014;349:g4287 doi: 10.1136/bmj.g4287 (Published 3 July 2014)

ENDGAMES

STATISTICAL QUESTION

Pitfalls of statistical hypothesis testing: type I and type II errors

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 the effects of a multidimensional lifestyle intervention on aerobic fitness and adiposity in predominantly migrant preschool children. A cluster randomised controlled trial study design was used. Intervention included a physical activity programme, plus lessons on nutrition, media use (use of television and computers), and sleep and adaptation of the built environment of the preschool class. The study lasted from August 2008 to June 2009. The control group did not receive any intervention and continued their regular school curriculum, which included one 45 minute physical activity lesson a week in the gym.[1]

In total, 40 preschool classes were recruited in areas with a high migrant population in the German and French speaking regions of Switzerland. Classes were randomised to intervention or control after stratification for linguistic region. The primary outcomes were aerobic fitness as measured by the 20 minute shuttle run test and body mass index (BMI). Secondary outcomes included motor agility, percentage body fat, and waist circumference.

The critical level of significance for statistical testing was set at 0.05 (5%). Intervention resulted in significantly increased aerobic fitness compared to control (intervention minus control) (mean difference: 0.32 stages; 95% confidence interval 0.07 to 0.57; P=0.01), but no difference in BMI (−0.07 kg/m^2, −0.19 to 0.06; P=0.31). The authors concluded that a multidimensional intervention increased aerobic fitness and reduced body fat but not BMI in predominantly migrant preschool children.

Which of the following statements, if any, are true?

a) If in the population there was no difference between intervention and control in mean aerobic fitness, then a type I error occurred for the statistical test of aerobic fitness.

b) If in the population there was a difference between intervention and control in BMI, then a type II error occurred for the statistical test of BMI.

c) The probability of a type I or type II error occurring would be reduced by increasing the sample size.

Answers

Statements *a*, *b,* and *c* are all true.

The aim of the trial was to establish the effects of a multidimensional lifestyle intervention on aerobic fitness and adiposity in predominantly migrant preschool children. The intervention was compared against control that consisted of no intervention, with children continuing their regular school curriculum. The treatment groups were compared with regard to primary endpoints using traditional statistical hypothesis testing, described in a previous question.[2]

Statistical hypothesis testing quantifies the evidence that the collected data support a specified hypothesis about the population. Traditional statistical hypothesis testing starts at the position of equipoise as specified by the null hypothesis. For the trial above, the null hypothesis for the primary outcome of aerobic fitness states that, in the population of predominantly migrant preschool children from which the sample was obtained, no difference exists between intervention and control in mean aerobic fitness as measured by the 20 minute shuttle run test. The aim was to establish whether the sample data supported this position or provided evidence of a difference, as specified by the alternative hypothesis. The alternative hypothesis states that a difference exists, and that in the population sampled, the mean aerobic fitness for those receiving intervention is not the same as for those in the control group. No direction is specified—the alternative hypothesis is two sided—intervention could be inferior or superior to control in mean aerobic fitness (as measured by the 20 minute shuttle run test). The statistical null and alternative hypotheses would have been stated before the trial started, if only conceptually. The statistical hypotheses for the primary outcome of BMI would have been constructed in a similar fashion.

The P value resulting from a statistical hypothesis test is used to establish whether the sample data support the null hypothesis, or provide evidence of a difference as specified by the alternative hypothesis. The P value is a probability and is derived using the sample data. It represents the strength of evidence in support

p.sedgwick@sgul.ac.uk

of the null hypothesis. A large P value suggests that the sample data support the null hypothesis, whereas a small P value suggests they do not. Conventionally the critical level of significance is set at 0.05 (5%). The P value for the statistical test of aerobic fitness was P=0.01, which is less than 0.05. Therefore, there was little evidence to support the null hypothesis, and it was rejected in favour of the alternative hypothesis. There was a statistically significant difference in aerobic fitness at the 0.05 level of significance—observation of the sample data shows that intervention increased aerobic fitness compared to control. The P value for the statistical test of BMI was P=0.31, which is greater than 0.05. Therefore, there was no evidence to reject the null hypothesis in favour of the alternative. Hence there was not a statistically significant difference in BMI at the 0.05 level of significance between intervention and control. It was inferred that intervention did not reduce BMI.

Statistical hypothesis testing makes inferences on the basis of a single sample. The sample in the trial above was just one of a potentially infinite number that could have been taken from the population of predominantly migrant preschool children. A further study, with a different sample, may have given different results. Obviously it is hoped the inferences in the trial above are representative of what would be observed in the population—that is, the sample estimates are similar in magnitude to the population parameters. A population parameter is conceptual and is the difference between intervention and control in, for example, body mass index that would be observed if intervention and control were administered to the entire population. Any difference between the sample estimate and population parameter would be due to sampling error, described in a previous question.[3]

Two types of error can be made in statistical hypothesis testing—type I and type II errors. In the trial above, a type I error would have occurred if the trial provided evidence of a significant difference and the null hypothesis was rejected in favour of the alternative, when in fact there was no difference in outcome in the population. There was a statistically significant difference between intervention and control in aerobic fitness (P=0.01). There was sufficient evidence to reject the null hypothesis in favour of the alternative, with the inference that intervention results in increased aerobic fitness. However, if in the population there was no difference between intervention and control in mean aerobic fitness, then an incorrect inference would have been made based on the statistical testing and a type I error would have occurred (*a* is true).

A type II error would have occurred if the study provided no evidence to reject the null hypothesis in favour of the alternative, when in fact there was a difference in outcome in the population. Although the intervention group had a lower mean BMI than control following intervention, the difference was not statistically significant (mean difference: −0.07 kg/m^2, 95% CI: −0.19 to 0.06; P=0.31). Therefore, there was insufficient evidence to reject the null hypothesis in favour of the alternative, with the inference that intervention does not result in decreased BMI. However, if in the population there was a difference between intervention and control in mean BMI, then an incorrect inference would have been made based on the statistical testing and a type II error would have occurred (*b* is true).

Type I errors and type II errors are conceptual and it is not possible to ascertain if they have occurred after statistical testing. However, it is possible to limit the probability of type I errors and type II errors occurring. The probability of making such errors is principally influenced by sample size.[4] As sample size increases and approaches that of the population, sample estimates will become similar in magnitude to the population parameters, making it less likely that a type I error or type II error will occur (*c* is true).

Competing interests: None declared.

1 Puder JJ, Marques-Vidal P, Schindler C, Zahner L, Niederer I, Bürgi F, et al. Effect of multidimensional lifestyle intervention on fitness and adiposity in predominantly migrant preschool children (Ballabeina): cluster randomised controlled trial. *BMJ* 2011;343:d6195.
2 Sedgwick P. Understanding statistical hypothesis testing. *BMJ* 2014;348:g3557.
3 Sedgwick P. What is sampling error? *BMJ* 2012;344:e4285.
4 Sedgwick P. Sample size: how many participants are needed in a trial? *BMJ* 2013;346:f1041.

Cite this as: *BMJ* 2014;349:g4287