STATISTICAL QUESTION

Understanding P values

Philip Sedgwick *reader in medical statistics and medical education*

Institute of Medical and Biomedical Education, St George's, University of London, London, UK

Researchers investigated whether a low glycaemic index diet in pregnancy reduced the incidence of macrosomic (large for gestational age) infants in an at risk group. A randomised controlled trial study design was used. Participants were 800 women without diabetes, all in their second pregnancy, who had previously delivered an infant weighing more than 4000 g. The intervention consisted of a low glycaemic index diet from early pregnancy. The control treatment was no dietary intervention. The primary outcome was birth weight.[1]

Treatment groups were compared in mean birth weight using the independent samples *t* test. Hypothesis testing was two tailed, with a critical level of significance of 0.05 (5%). The mean birth weight in the intervention group was greater than in the control group, although the difference was not significant (4034 g (standard deviation 510) *v* 4006 (497); mean difference 28.6 g; 95% confidence interval −45.6 to 102.8; P=0.449). The researchers concluded that a low glycaemic index diet in pregnancy did not reduce the incidence of large for gestational age infants in a group at risk of fetal macrosomia.

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

a) Statistical hypothesis testing based on a critical level of significance is a dichotomous test

b) The P value provides a direct statement about the direction of a difference between treatment groups in mean birth weight

c) The P value is the probability that the alternative hypothesis was true

Answers

Statement *a* is true, whereas *b* and *c* are false.

The aim of the trial was to investigate the effects of a low glycaemic index diet in pregnancy. The treatment groups were compared in the primary outcome of birth weight using traditional statistical hypothesis testing, described in a previous question.[2] The sample estimate of the population parameter of the difference between the intervention and control in mean birth weight was 28.6 g. The purpose of the hypothesis test was to establish whether the difference in mean birth weight seen in the trial also existed in the population.

Statistical hypothesis testing involves the statement of the null and alternative hypotheses. Traditional statistical hypothesis testing starts at the position of equipoise as specified by the null hypothesis. For the primary outcome in the trial above, the null hypothesis states that in the population of women at high risk of delivering a macrosomic infant from which the sample was obtained, no difference exists between the intervention and control treatments in mean birth weight. The alternative hypothesis states that a difference exists—that is, in the population sampled a difference exists between treatments in mean birth weight. No direction is specified—the alternative hypothesis is two sided—the mean birth weight for the low glycaemic index diet group could be greater or smaller than for the control group. The aim was to establish whether the sample data supported the null hypothesis or provided evidence of a difference between treatment groups as specified by the alternative hypothesis. A statistical hypothesis test involving a two sided alternative hypothesis is sometimes referred to as two tailed.

The P value for the statistical test of the primary outcome of birth weight was P=0.449. It is a probability and was derived using the sample data. In this instance the P value resulted from the statistical test known as the independent samples *t* test.[3] The P value represents the strength of evidence in support of the null hypothesis. A large P value suggests that the sample data support the null hypothesis, whereas a small P value suggests that they do not. The cut off between a large and a small P value is conventionally set at 0.05 (5%), which is termed the critical level of significance. If the value of P is 0.05 or more, the sample data have provided insufficient evidence to reject the null hypothesis, whereas if it is less than 0.05 (5%), the evidence is sufficient to reject the null hypothesis in favour of the alternative. As such, statistical hypothesis testing based on a critical level of significance is a dichotomous test (*a* is true). It is recommended that the P value is always used to report the results of a statistical hypothesis test, rather than "not significant (NS)" or "significant (S)," because it provides a continuous measure of the strength of evidence in support of the null hypothesis.

p.sedgwick@sgul.ac.uk

The derivation of the P value for the hypothesis test of birth weight was based on the theoretical situation of sampling an infinite number of times. This can be demonstrated using complex theory, although details of these are beyond the scope of this article. Sampling would be from the population of women at high risk of delivering a macrosomic infant, in which it is assumed there was no difference between the intervention and control treatments in mean birth weight. Each of the infinite number of samples would be the same size and would be obtained under the same conditions as for the study above. The sample size for the trial above was determined a priori. The smallest effect of clinical interest was a difference of 102 g between treatment groups in mean birth weight—that is, for the effects of treatment to be considered clinically significant a difference of 102 g or more was needed. To observe this difference with 90% power and a critical significance level of 5%, 360 participants were needed in each treatment group. Sample size calculations for clinical trials have been described in a previous question.[4]

The infinite number of samples would be selected at random from the population and therefore not consist of the same population members. Hence, each sample would provide a different estimate of the population parameter of the difference between treatments in mean birth weight. For these samples, the mean birth weight could be higher or lower for the intervention group than for the control group. Because the critical level of significance was set at 0.05 (5%), the null hypothesis would be rejected in favour of the alternative for 5% of these infinite number of samples. In particular, these 5% of samples would be those that demonstrated the smallest effect of clinical interest—that is, a difference of 102 g or more between treatment groups in birth weight. This would be regardless of whether the mean birth weight for the intervention group was greater or smaller than for the control group—the direction of the difference is ignored. Therefore, it is for these 5% of samples with the largest difference in mean birth weight that a statistically significant difference in outcome would exist between treatment groups.

The P value for the statistical test of birth weight was P=0.449. The P value represents the proportion of the theoretical infinite number of samples—that is, 0.449 (49.9%)—that have a mean difference in birth weight equal to, or greater than, that observed in the trial above. This is irrespective of whether the mean birth weight was higher or lower for the intervention group than for the control group. More formally, the P value is the probability of obtaining the observed difference between treatment groups in mean birth weight (or a larger one), irrespective of the direction, if there was no difference between treatment groups in mean birth weight in the population, as specified by the null hypothesis. The P value for the statistical test of the primary outcome of birth weight was P=0.449, which was larger than the critical level of significance (0.05). Hence there was no evidence to reject the null hypothesis in favour of the alternative. The inference is that there was no evidence that the intervention and control treatments differed in mean birth weight in the population.

The P value alone cannot provide any direct statement about the size of the difference between treatment groups in mean birth weight. Furthermore, the P value does not provide any indication of the direction of the difference between treatment groups—that is, whether the mean birth weight was higher or lower for the intervention group than for the control group (*b* is false). The 95% confidence interval for the difference between treatment groups in mean birth weight was therefore presented, because it enabled a statement to be made about the size and direction of the difference between treatment groups.

The P value does not seem to be well understood. This may be because it is an abstract concept. Although the derivation of the P value is based on the theoretical concept of sampling from the population an infinite number of times, in practice a single sample is obtained. The P value is often misinterpreted—for example, it is often thought that the P value is the probability that the null hypothesis, or the alternative hypothesis, is true or false (*c* is false). As described above, the P value indicates whether the sample data support the null hypothesis or lend support to the alternative. This distinction is important, because theoretically it would be difficult to prove that a hypothesis is true or false. The null or alternative hypothesis may be true for a population. However, the only way to prove or disprove a statistical hypothesis is to sample the entire population, which is not feasible. The sample for a study is one of a theoretical infinite number taken from a population and is therefore prone to sampling error.[5] Small samples are more likely to result in a type I or II error when hypothesis testing. Such errors have been described in a previous question.[6] Clinical trials sometimes recruit too many participants and are overpowered. In this situation, a difference between treatment groups in outcome that is not clinically significant may be found to be statistically significant.[7 8] Therefore, it may be incorrect, and hence misleading, to infer that the null or alternative hypothesis is true or false from the results for a statistical hypothesis test for a single sample.

Competing interests: None declared.

1 Walsh JM, McGowan CA, Mahony R, Foley ME, McAuliffe FM. Low glycaemic index diet in pregnancy to prevent macrosomia (ROLO study): randomised control trial. *BMJ* 2012;345:e5605.
2 Sedgwick P. Understanding statistical hypothesis testing. *BMJ* 2014;348:g3557.
3 Sedgwick P. Independent samples t test. *BMJ* 2010;340:c2673.
4 Sedgwick P. Sample size: how many participants are needed in a trial? *BMJ* 2013;346:f1041.
5 Sedgwick P. What is sampling error? *BMJ* 2012;344:e4285.
6 Sedgwick P. Pitfalls of statistical hypothesis testing: type I and type II errors. *BMJ* 2014;349:g4287.
7 Sedgwick P. The importance of statistical power. *BMJ* 2013;347:f6282.
8 Sedgwick P. Clinical significance versus statistical significance. *BMJ* 2014;348:g2130.

Cite this as: *BMJ* 2014;349:g4550