BMJ 2015;350:h1113 doi: 10.1136/bmj.h1113 (Published 27 February 2015)

STATISTICAL QUESTION

Confidence intervals, P values, and statistical significance

Philip Sedgwick *reader in medical statistics and medical education*

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

The efficacy of nicotine patches in pregnant women who smoked was investigated using a randomised placebo controlled trial. The intervention was the administration of 16 hour nicotine patches until the time of delivery. Participants were pregnant women over 18 years who smoked at least five cigarettes a day and whose babies were between 12 and 20 weeks' gestation. In total, 402 women were recruited from 23 maternity wards throughout France. Participants were randomised to the intervention (n=203) or placebo patches (n=199).[1]

The outcome measures included achievement of complete abstinence until delivery and birth weight. Complete abstinence was achieved by 5.5% (n=11) of women in the nicotine patch group and 5.1% (n=10) in the placebo group (odds ratio 1.08, 95% confidence interval 0.45 to 2.60). The mean birth weight was higher in the nicotine patch group (3065 (standard error 44 g) *v* 3015 g (44 g); difference 50 g, −71.1 to 172.3).

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

a) The odds ratio for abstinence until delivery was statistically significant at the 5% level because the associated 95% confidence interval did not straddle zero

b) The difference between treatment groups in mean birth weight was not statistically significant at the 5% level because the associated 95% confidence interval straddled zero

c) A 95% confidence interval provides a test of the statistical hypotheses at the 5% level of significance.

Answers

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

The odds ratio comparing the intervention group with the placebo group in smoking abstinence until delivery was 1.08 (0.45 to 2.60), whereas mean birth weight was greater for the intervention group (difference 50 g, −71.1 to 172.3). The 95% confidence intervals are interval estimates for the population parameters of the odds ratio of smoking abstinence until delivery and the difference in mean birth weight. The confidence intervals represent the uncertainty of the sample in estimating the population parameters owing to sampling error.[2] For each outcome measure, a statistical hypothesis test could have been undertaken to establish if there was a significant difference between treatment groups. Statistical hypothesis testing has been described in a previous question.[3] The critical level of significance when hypothesis testing is typically set at 0.05 (5%).[4] There is a unique association between a 95% confidence interval for the population parameter and the 5% level of significance when hypothesis testing.

Traditional statistical hypothesis testing was used to establish whether the difference between treatment groups in the proportion of women who achieved abstinence until delivery was significant. Treatment groups were compared using an odds ratio, which has been described in a previous question.[5] Hypothesis testing started at the position of equipoise. The null hypothesis stated that in the population of pregnant women from which the sample was obtained, there was no difference between treatment groups in the odds of abstinence—that is, the odds ratio equalled 1.0 (unity). The alternative hypothesis was two sided—that is, the odds ratio was lower or higher than unity. The critical level of significance was 0.05 (5%). The researchers reported that the P value for the statistical test of the odds ratio was 0.87, and that therefore the difference between treatment groups in abstinence until delivery was not significant at the 5% level. The inference of statistical significance could have been made on the basis of the 95% confidence interval. The 95% confidence interval for the population odds ratio was 0.45 to 2.60. Because the 95% confidence interval straddled unity, it can be inferred that the difference between treatment groups in the achievement of complete abstinence until delivery was not significant at the 5% level—that is P>0.05 (*a* is false).

Generally, if the 95% confidence interval for a ratio statistic, such as a relative risk, hazard ratio, or odds ratio, straddles unity, then the test of the statistical hypotheses for the comparison of groups in the outcome will not be significant at the 5% level. If the 95% confidence interval excludes unity then the test of the statistical hypotheses will be significant at the 5% level, and the null hypothesis will be rejected in favour of the alternative.

p.sedgwick@sgul.ac.uk

If one of the limits of a 95% confidence interval is equal to 1.0 (unity), then the P value will be equal to 0.05 (5%). The 95% confidence interval for a ratio statistic will never straddle zero—the lower limit will be above zero with the upper limit bounded by infinity. The inference of statistical significance at the 5% level based on a 95% confidence interval for a ratio statistic is centred around whether the confidence interval straddles unity. This process of inferring statistical significance for a ratio statistic should not be confused with that for the difference between treatment groups in an outcome, such as birth weight, which as described below is based on whether the confidence interval straddles zero (*a* is false).

Traditional hypothesis testing was also used to test the difference between treatment groups in mean birth weight. The null hypothesis would have stated that in the population of pregnant women from where the sample was obtained, there was no difference between treatment groups in mean birth weight—that is, the mean difference was zero. The alternative hypothesis was two sided—that is, the mean difference was lower or higher than zero. The critical level of significance was 0.05 (5%). The researchers reported that the P value for the statistical test of the mean difference in birth weight was 0.41, and that therefore the difference between treatment groups in mean birth weight was not significant at the 5% level. The inference of statistical significance could have been made on the basis of the 95% confidence interval. The 95% confidence interval for the population difference in mean birth weight was (−71.1 to 172.3 g). Because the 95% confidence interval straddled zero, it can be inferred that the difference between treatment groups in mean birth weight was not significant at the 5% level—that is P>0.05 (*b* is true).

Generally, if the 95% confidence interval for the difference in an outcome variable between two treatment groups, such as the difference in means or percentages, straddles zero then the test of the statistical hypotheses for the difference will not be significant at the 5% level. If the 95% confidence interval excludes zero then the test of the statistical hypotheses will be significant at the 5% level, and the null hypothesis will be rejected in favour of the alternative. If one of the limits of a 95% confidence interval is equal to zero, then the P value will be equal to 0.05 (5%).

Statistical hypothesis testing that is based on the critical level of significance of 5% as described above is a dichotomous test. The derived P value is a probability and a measure of the strength of the evidence provided by the sample in support of the null hypothesis. If the P value is less than 0.05 (5%), the null hypothesis is rejected in favour of the alternative. Generally, a smaller P value indicates a larger difference or stronger association between treatment groups in the outcome. However, the P value alone cannot inform any direct statement about the size of the treatment effect. As described above, the 95% confidence interval is similar to the P value in that it can also be used to perform a dichotomous statistical hypothesis test at the 5% level of significance (*c* is true). However, in contrast to the P value, the 95% confidence interval indicates the direction and size of the treatment effect. In particular, a 95% confidence interval permits the clinical significance of a treatment effect to be evaluated in addition to the statistical significance.[6] Therefore, confidence intervals are preferable to P values. Nonetheless, it is useful for both confidence intervals and P values to be reported in journal articles because they provide statistical measures that complement each other.

Competing interests: None declared.

1 Berlin I, Grangé G, Jacob N, Tanguy M-L. Nicotine patches in pregnant smokers: randomised, placebo controlled, multicentre trial of efficacy. *BMJ* 2014;348:g1622.
2 Sedgwick P. Understanding confidence intervals. *BM* J 2014;349:g6051.
3 Sedgwick P. Understanding statistical hypothesis testing. *BMJ* 2014;348:g3557.
4 Sedgwick P. Understanding P values. *BMJ* 2014;349:g4550.
5 Sedgwick P. Odds and odds ratios. *BMJ* 2013;347:f5067.
6 Sedgwick P. Clinical significance versus statistical significance. *BMJ* 2014;348:g2130.

Cite this as: *BMJ* 2015;350:h1113