Hypothesis testing and p values: how to interpret results and reach the right conclusions

Allison Shorten, Brett Shorten

10.1136/eb-2013-101255

Yale University School of Nursing, New Haven, Connecticut, USA

Correspondence to:
Dr Allison Shorten
Yale University School of Nursing, 100 Church Street South, PO Box 9740, New Haven, CT 06536 USA;
allison.shorten@yale.edu

Whenever we encounter a research finding based on the interpretation of a p value from a statistical test, whether we realise it or not, we are discussing the result of a formal hypothesis test. This is true irrespective of whether the test involves comparisons of means, Odds Ratios (ORs), regression results or other types of statistical tests. As readers of research, it is important to understand the underlying principles of hypothesis testing, so that when faced with statistical results, we reach the right conclusions and make good decisions about which findings are robust enough to be translated into clinical practice.

The article by Yinon *et al*[1] featured in a recent EBN commentary, will be used to illustrate four simple steps involved in hypothesis testing.[2] The authors of this paper explored the possible benefits of antenatal steroid administration in the context of late preterm birth (>34 weeks gestation). One of the key outcomes of interest included the incidence of babies being admitted to a special care unit (SCU). It was hypothesised that steroid administration would lead to better respiratory function and therefore reduction in SCU admissions. In the sample, 14 of 83 neonates (almost 17%) in the experimental (steroid) group were admitted to SCU, compared with 24 of 84 neonates (almost 29%) in the control (no steroids) group.[1] At first glance we see a difference in the two groups, however, we need to look further and decide whether the differences found represent real differences in SCU admission rates due to antenatal steroid administration. It may be plausible that the differences observed are due to random differences within the sample studied. Let's follow four simple steps to reach a conclusion about these results.

Step 1

Identify both a *null* and an *alternative hypothesis*. As the name implies, the null hypothesis is that there are no differences between the two groups. In this case SCU admission rates would be the same whether steroids were administered or not. The alternative hypothesis would be that there is a difference in SCU admission rates between the two groups.

Step 2

Identify the *test statistic* used to test the hypothesis. In this case, the researchers used the χ^2 (Chi-Square) statistic and calculated a p value of 0.07. Table 1 provides the information needed to calculate this. There are online χ^2 calculators available to check the result for yourself.[3]

Table 1 Information used to calculate the χ^2

Outcome of interest	Steroids used	No steroids used
Admitted to SCU	14	24
Not admitted to SCU	69	60

χ^2=3.25, with 1 degree of freedom.
SCU, special care unit.

At this point, it is important to pause and imagine what might happen if we could perform this same study thousands of times by selecting many different samples of 83 women to whom steroids would be given and 84 women as a control group, observing SCU admissions for each group. If the null hypothesis is true, that is, SCU admission is equally likely for both groups and there is no benefit in steroid administration, the 'true' χ^2 value would be 1. Therefore, on average the χ^2 values calculated for the thousands of samples would equal 1. Sometimes, due to random sampling variability, the χ^2 would be somewhat higher than 1 and sometimes lower than 1.

Step 3

Calculate the p value and decide whether the value of 3.25 is sufficiently higher than 1 to convince us that SCU admission rates do in fact differ between the two groups. The authors reported a p value of 0.07 which indicates that, if we performed this study thousands of times, and if the null hypothesis is in fact true, we would expect 0.07, or 7%, of χ^2 values to be at least as extreme (greater than 1) as the value of 3.25. Therefore, if we were to reject the null hypothesis in favour of the alternative and conclude that steroid use actually reduces SCU admission rates, there would be a 7% chance that we would be incorrect in doing so. When we set our critical p value level (α) at 0.05, we are stating that we are willing to risk only a 5% chance of error when we reject the null hypothesis.[4]

Step 4

Accept or reject the null hypothesis. In this case accept (p=0.07 is greater than 0.05). It is critical at this point to realise that we have not proven the null hypothesis to be correct. We cannot state that *there was no difference in the rate of special care unit admissions* merely because the p value is 0.07. Specifically, we have not demonstrated that the null hypothesis is true, but have decided that the evidence is not robust enough to disprove it. If only 17% of neonates were admitted to SCU when steroids were used, compared to 29% for the control group, it would clearly be incorrect to state that we have in any sense 'proven' that the null hypothesis is true. The correct conclusion is that we must default to what our old statistics professor used to term 'our original state of ignorance'—that is, we still do not know whether steroid administration affects SCU admission rates and further research is required.

Competing interests None.

Editor's choice
Scan to access more free content

References

1. **Yinon Y,** Haas J, Mazaki-Tovi S, *et al*. Should patients with documented fetal lung immaturity after 34 weeks of gestation be treated with steroids? *Am J Obstet Gynecol* 2012;**207**:222.e1–4.
2. **Kamath-Rayne B.** Cohort study finds newborn respiratory complications less common when mothers of babies with foetal lung immaturity at 34 to 37 weeks gestation given antenatal steroids. *Evid Based Nurs* Published Online First 13 Feb 2013. doi: 10.1136/eb-2012-101105
3. **Preacher KJ.** Calculation for the chi-square test: an interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. 2001. http://quantpsy.org (accessed 28 Jan 2013).
4. **Forbes DA.** What is a p value and what does it mean? *Evid Based Nurs* 2012;**15**:34.