ANESTHESIA & ANALGESIA® Statistical Minute

Related Article, see p 1312

# Chi-square Tests in Medical Research

Patrick Schober, MD, PhD, MMedStat,* and Thomas R. Vetter, MD, MPH†

| | | | Treatment Group | | |
|---|---|---|---|---|---|
| | | | Phenylephrine | Norepinephrine | Total |
| Bradycardia | Yes | Observed Number | 21 | 6 | 27 |
| | | *Expected Number* | *13.5* | *13.5* | |
| | No | Observed Number | 35 | 50 | 85 |
| | | *Expected Number* | *42.5* | *42.5* | |
| | Total | | 56 | 56 | 112 |

**Figure.** Contingency table with data from Sharkey et al[1] showing the observed and expected counts (number of patients) with and without bradycardia per treatment group. Assuming that the probability of developing bradycardia is independent of the group allocation (null hypothesis), 13.5 patients with bradycardia would be expected in each group (given a total of 27 patients who developed bradycardia and equal sample size in both groups). Pearson $\chi^2$ test compares observed to expected frequencies.

**KEY POINT:** A $\chi^2$ test is commonly used to analyze categorical data, but valid statistical inferences rely on its test assumptions being met.

In this issue of *Anesthesia & Analgesia*, Sharkey et al[1] report a randomized trial comparing the incidence of bradycardia after phenylephrine versus norepinephrine to prevent and treat spinal-induced hypotension in women undergoing cesarean delivery with spinal anesthesia. The authors used a chi-square ($\chi^2$) test to compare the groups and observed a lower incidence of bradycardia in the norepinephrine group.

A $\chi^2$ test commonly either compares the distribution of a categorical variable to a hypothetical distribution or tests whether 2 categorical variables are independent. We focus here on the Pearson $\chi^2$ test of independence used by Sharkey et al.[1]

The Pearson $\chi^2$ test evaluates the null hypothesis that 2 categorical variables (eg, treatment group [norepinephrine versus phenylephrine] and outcome [bradycardia versus no bradycardia]) are not associated with each other.[2] In the study by Sharkey et al,[1] a total of 27/112 (24.1%) patients developed bradycardia (Figure). Assuming independence between treatment and bradycardia, the same percentage would be expected in each group (ie, 13.5 patients with bradycardia per group). The $\chi^2$ test then compares the observed to expected frequencies. The reported *P* value of .001 suggests that it is very unlikely to observe a difference this large or larger if the null hypothesis was true,[3] supporting the conclusion that there is an association between treatment and outcome. While in this example, both categorical variables have 2 levels (2 groups, 2 outcomes), a $\chi^2$ test can more generally be used when variables have multiple categories.

However, valid inferences with a $\chi^2$ test rely on a number of assumptions,[2] including:

1. The actual frequencies can be crosstabulated in a contingency table. It is not appropriate to use a $\chi^2$ test for percentages or other derived statistics.
2. The 2 variables are nominal—there is no natural ordering of the categories.
3. The observations are independent.
4. The expected count or frequency is ≥5 in more than 75%–80% of the cells in the contingency table, and there is no expected cell count of 0.

When analyzing ordinal data, the Pearson $\chi^2$ test ignores the order (assumption #2), and the Mantel–Haenszel $\chi^2$ test provides more power to test for an association.[2] Repeated measurements in the same subjects are a common violation of assumption #3. Such data require tests that account for the pairing (eg, McNemar test) or longitudinal models that account for the within-subject correlation.[4] When expected counts are lower than specified in assumption #4, Fisher exact test can be used.[2]

Importantly, the $\chi^2$ test assesses for an association but does not provide information on the strength of the association or on whether the relationship is causal. While a properly conducted randomized controlled trial allows causal inferences, observed associations can be confounded in uncontrolled studies. In such cases, techniques that control for confounding, such as multivariable logistic regression,[5] are strongly preferred.

## REFERENCES

1. Sharkey AM, Siddiqui N, Downey K, Ye XY, Guevara J, Carvalho JCA. Comparison of intermittent intravenous boluses of phenylephrine and norepinephrine to prevent and treat spinal-induced hypotension in cesarean deliveries: randomized controlled trial. *Anesth Analg*. 2019;129:1312–1318.
2. Vetter TR, Mascha EJ. Unadjusted bivariate two-group comparisons: when simpler is better. *Anesth Analg*. 2018;126:338–342.
3. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: what do P values and confidence intervals really represent? *Anesth Analg*. 2018;126:1068–1072.
4. Schober P, Vetter TR. Repeated measures designs and analysis of longitudinal data: if at first you do not succeed-try, try again. *Anesth Analg*. 2018;127:569–575.
5. Vetter TR, Schober P. Regression: the apple does not fall far from the tree. *Anesth Analg*. 2018;127:277–283.

From the *Department of Anesthesiology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; and †Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Austin, Texas.

Address correspondence to Patrick Schober, MD, PhD, MMedStat, Department of Anesthesiology, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands. Address e-mail to p.schober@amsterdamumc.nl.