EDITORIAL

JORNAL VASCULAR BRASILEIRO
ISSN 1677-7301 (Online)

Comparing categorical variables in clinical and experimental studies

Comparação entre variáveis categóricas em estudos clínicos e experimentais

Anna Carolina Miola[1], Hélio Amante Miot[1]

How to cite: Miola AC, Miot HA. Comparing categorical variables in clinical and experimental studies. J Vasc Bras. 2022;21:e20210225. https://doi.org/10.1590/1677-5449.20210225

Many studies of a quantitative nature use qualitative variables, within both the biomedical sciences and the social sciences. These variables are also known as categorical and their magnitude is expressed in terms of the frequency with which each of their categories occurs. Qualitative variables can be subdivided into dichotomous (for example, sex, death, cure), ordinal (for example, cancer staging, pulse amplitude, functional class, phototype, anesthetic risk) or polytomous/multinomial (for example, sexual orientation, ABO type, marital status, religion, race, aneurysm type, type of chronic ulcer).[1-3]

When qualitative variables are employed, the phenomenon measured can be represented as the percentage of occurrence in each category, and subgroups should be compared in terms of the proportion of the sample that is attributed to each class.[3] There is an extensive literature on techniques for statistical analysis of qualitative variables;[4-6] whereas this text will deal with comparison of proportions between categorical variables. Comparative analysis of proportions between subgroups employs different concepts from parametric statistics, providing lower statistical power (larger type II error) in analogous situations, such as when a quantitative variable (for example, age) is categorized (for example, < 30 years, 30–59 years, ≥ 60 years).[7,8]

According to frequentist statistics,[9] the probability of a proportion of events selected at random, without replacement of cases, can be generalized from the chi-square distribution, while Pearson's chi-square test is based on the difference between the frequencies observed and the frequencies ideally expected for each category and can be used to compare how well a sample fits a known distribution (for example, for comparison with the literature) or independence between different samples.[10] Despite the popularity of Pearson's chi-square test, other methods such as the G test (likelihood ratio) and the Goodman test (contrasts between proportions) are also used to compare proportions. However, absolute superiority between them has not yet been systematically defined.[11-14]

An observed proportion's fit can be compared to a description from the literature or a theoretical prediction (for example, expression of a phenotype according to segregation of a gene).[15] For example, Tamega et al.[16] studied ABO and Rh blood typing of 69 patients with lupus erythematosus, comparing them against the expected frequencies of these categories among blood donors at the institution. Pearson's chi-square test (of fit) returned a *p*-value 0.081 for ABO types and a *p*-value of 0.721 for Rh types, accepting the hypothesis that these blood type classes did not differ from what was expected in the local population.[16]

In clinical-epidemiological research, it is highly usual to present an initial descriptive table containing demographic data on subgroups, in order to demonstrate their homogeneity. For example, Amiri et al.[17] included 110 cases and 110 controls in a cross-sectional study to test associations between anthropometric indices and type 2 diabetes mellitus. Of these diabetic patients, 75 (51%) were female, whereas there were only 72 women (49%) in the control group. In this sample, the difference in proportion between the groups (2%) was not considered significant (*p*-value = 0.668) for this dichotomous variable according to Pearson's chi-square test (of independence).

While versatile, Pearson's chi-square test has inadequate performance (larger type I error) in smaller samples ($n \leq 40$), especially in which > 20% of

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Faculdade de Ciências Médicas e Biológicas de Botucatu, Botucatu, SP, Brasil.
Financial support: None.
Conflicts of interest: No conflicts of interest declared concerning the publication of this article.
Submitted: December 11, 2021. Accepted: January 20, 2022.

The study was carried out at Departamento de Dermatologia, Faculdade de Ciências Médicas e Biológicas de Botucatu, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Botucatu, SP, Brazil.

expected values are ≤ 5, which is relatively common in biomedical research scenarios. Several procedures are recommended in this situation, ranging from combining categories to increase the predicted value (for example, dichotomizing skin colors as white vs. not white, combining less common B blood groups with AB) or using other statistical tests.

There is an intense academic debate about which analytical strategies should be adopted for situations in which Pearson's chi-square test is contraindicated, while different tests for categorical variables can behave differently depending on the manner in which the data are collected (randomized or not), since a large proportion of studies do not employ a completely randomized sampling structure.[18-20] The Barnard and Boschloo exact tests are two examples that correct for these limitations for 2 × 2 contingency tables.[21,22] In turn, the G test (with Williams' correction) can be used for multinomial comparisons in situations in which Pearson's chi-square test is contraindicated.[21,23] Estimates of the (exact) *p*-value using resampling (bootstrapping) or Monte Carlo simulation are also effective in cases with modest samples or subgroups with a low predicted rate of occurrence.[19,24]

Fisher's exact test is cited in many texts as a solution for cases in which Pearson's chi-square test is not indicated, but it inflates the type II error, in addition to being based on a conditional probability model, which contrasts with what is usually proposed in biomedical research biomedical research (variable marginal totals).[25,26] Along the same lines, correcting Pearson's chi-square test with Yates' procedure is excessively conservative in 2 × 2 tables. Use and interpretation of these tests should be parsimonious when they return *p*-values close to the significance level.[22,24]

For more complex designs, involving interaction between more than two categorical variables or multivariate adjustments in which the dependent variable is categorical, other methods of analysis can be used, such as Poisson regression (log-linear), logistic regression, and multinominal regression, which, as with Pearson's chi-square test, are penalized in cases with low frequencies in subgroups. On the other hand, multivariate methods, such as multiple correspondence analysis, are unaffected by the contingencies of tests of hypotheses and can support exploratory analyses of categorical data.[4,27,28] Meanwhile, the problems linked to analysis of ordinal data and calculation of sample sizes for studies involving proportions have been covered previously.[2,29-32]

When the results of comparisons of multinomial variables are significant, it remains to be determined which of the internal proportions diverge from the expected, since the test result (for Pearson's chi-square test, for example) refers to the overall behavior of the proportions, so it is necessary to proceed to a *post hoc* analysis of the subcategories. Analysis of the residuals of the contingency table (standardized and adjusted) is a widely-used strategy that returns a *Z* statistic (Zres) for each proportion found, enabling multiple comparison between them to identify which specific variables most contribute to the result observed in the global test.[33] By analyzing the residuals shown in Table 1, it can be concluded that cancer patients referred from clinics exhibited more incidental tomographic diagnoses of pulmonary thromboembolism than those admitted via the emergency room, however, no differences were found in the proportions from inpatients and those form ICU.[34]

Another option for analysis of subgroups is Goodman and Kruskal's *lambda* test, which is a measure of the proportional reduction in error in the contingency table analysis for multinomial data, indicating the point to which modal categories and frequencies for each value of the independent variable differ from the values of the independent variable.[35] In the same manner, the table can be partitioned into 2 × 2 subtables. However, the multiple comparisons must be adjusted to control inflation of the type I error, using the Bonferroni procedure, for example.[20]

Epidemiological research often employs dichotomous outcomes (for example, cure, death, sickness) to compare two or more groups (for example, placebo vs. treatment). Characteristics intrinsic to the designs of studies have led to a growing tendency for

Table 1. Analysis of residuals in data from Carneiro et al.[34] on the origin of cancer patients with pulmonary thromboembolism (PTE) on computed tomography of the thorax, when the finding was incidental or there was a prior suspicion.

Origin	No suspicion of PTE		PTE suspected previously	
	(*n* = 48)	Zres (*p*-value)	(*n* = 60)	Zres (*p*-value)
Clinic	28 (59%)	*+5.1 (<0.001)*	7 (12%)	*-5.1 (<0.001)*
Wards	16 (33%)	-0.2 (0.856)	21 (35%)	+0.2 (0.856)
Intensive care unit	2 (4%)	-1.4 (0.161)	7 (12%)	+1.4 (0.161)
Emergency room	2 (4%)	*-4.5 (<0.001)*	25 (43%)	*+4.5 (<0.001)*

p-value (global) < 0.001; Pearson's chi-square.

comparisons of these proportions to be estimated from their epidemiological measures of effect, such as odds ratios, relative risk, or prevalence ratios, rather than merely according to the results of statistical tests of proportion.[36,37] Both the *p*-value and the confidence interval of such associations can be calculated directly for these estimates using logistic, ordinal, multinominal, or Poisson regression models.[38]

The need to adjust results for covariates that are of importance in the causal model (for example, age, sex, smoking) has demanded wider adoption of these regression techniques for analysis of categorical data. Contingencies in the presence of modest samples or rarity of events in one of the categories can be overcome using bootstrapping techniques, resampling data more than 1,000 times. However, since these methods consider the relationships between subcategories, they do not deal adequately with cases in which one category is zero, in contrast with exact statistical techniques (Barnard's test, for example).

Table 2 shows examples of methods for analysis of comparisons between two hypothetical treatments (surgical vs. conventional) analyzed with tests of comparison of proportions and regression models, according to sample characteristics. In the special case of estimation of the magnitude of the effect of a study (for example, relative risk and odds ratios) in which there were zero occurrences of one of the categorical variables, it is possible to resort to the (artificial) addition of 0.5 units to the outcome of each group.[5,39,40]

Comparison of proportions between groups can also be evaluated unidirectionally or bidirectionally (one/two-tailed), since many analyses are by their nature one-directional, such as comparison of mortality rates from a disease among vaccinated and unvaccinated people or tests of non-inferiority between two treatments.[41] In such cases, the study hypothesis does not contemplate the possibility that the result could be analyzed bidirectionally, since there is only interest in the effect in one direction. One-tailed analyses of proportions do not enjoy consensus among epidemiologists because, although they have greater statistical power and require smaller samples, they increase the likelihood of type I error.[24] One-tailed analyses are widely used in studies of viability (pilot studies) and in proof-of-concept studies, which are conducted before traditional clinical trials.[42-44]

Situations that involve dependent data should be assessed with the McNemar test (2 × 2 tables), Cochran's Q test (several groups, dichotomous response), or generalized estimating equations. These analyses, in common with use of resampling techniques, one-tailed estimates, regressions and analyses of variables that demand multivariate adjustment, should be supervised by an experienced statistician.

Finally, comparison of categorical variables is a common need in biomedical studies and inferential conclusions can differ depending on the analytical method employed, especially when the frequencies in subgroups are low. The choice of analytical technique requires theoretical grounding and its description must be justified in the methodology in terms of the parameters for its use.

Table 2. Hypothetical examples of (two-tailed) comparisons of incidence of death from a disease treated with a surgical procedure or a conventional treatment.

Examples	Statistical test	Statistic/effect	*p*-value
2 deaths in 100 surgeries (2%)	Pearson's chi-square	χ^2 = 11.97; Df = 1	<0.001
vs.			
16 deaths in 100 conventional treatments (16%)	Poisson (robust) regression	RR = 0.13	0.005
		95%CI = 0.03 to 0.53	
1 death in 50 surgeries (2%)	Barnard	Score = 2.45	0.016
vs.			
8 deaths in 50 conventional treatments (16%)	Poisson regression	RR = 0.13	0.046
	(robust; 1000 resampling)	95%CI = 0.01 to 0.43	
Zero deaths in 50 surgeries (0%)	Barnard	Score = 2.95	0.004
vs.			
8 deaths in 50 conventional treatments (186%)		RR = 0.06[a]	0.034
		95%CI = 0.01 to 0.45	

Df = degrees of freedom; RR = relative risk; 95%CI = 95% confidence interval. [a]Relative risk calculated after addition of 0.5 units to the outcome in each group: 0.5 deaths among surgeries, 8.5 deaths among conventional treatments.

REFERENCES

1. Greenhalgh T. How to read a paper: statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ. 1997;315(7104):364-6. http://dx.doi.org/10.1136/bmj.315.7104.364. PMid:9270463.
2. Miot HA. Analysis of ordinal data in clinical and experimental studies. J Vasc Bras. 2020;19:e20200185. http://dx.doi.org/10.1590/1677-5449.200185. PMid:34211532.
3. Perkins SM. Statistical inference on categorical variables. Methods Mol Biol. 2007;404:73-88. http://dx.doi.org/10.1007/978-1-59745-530-5_5. PMid:18450046.
4. Pereira JCR. Análise de dados qualitativos: estratégias metodológicas para as ciências da saúde humanas e sociais. São Paulo: EdUSP; 1999.
5. Agresti A. An introduction to categorical data analysis. 2nd ed. New Jersey: John Wiley & Sons; 2020.
6. Quinn GP, Keough MJ. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press; 2002. http://dx.doi.org/10.1017/CBO9780511806384.
7. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41. http://dx.doi.org/10.1002/sim.2331. PMid:16217841.
8. Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol. 2011;32(3):437-40. http://dx.doi.org/10.3174/ajnr.A2425. PMid:21330400.
9. Zaslavsky BG. Bayesian versus frequentist hypotheses testing in clinical trials with dichotomous and countable outcomes. J Biopharm Stat. 2010;20(5):985-97. http://dx.doi.org/10.1080/10543401003619023. PMid:20721786.
10. Turner N. Chi-squared test. J Clin Nurs. 2000;9(1):93. PMid:11041649.
11. Goodman LA. On the multivariate analysis of three dichotomous variables. Ajs. 1965;71(3):290-301. http://dx.doi.org/10.1086/224088. PMid:5897475.
12. Eberhardt KR, Fligner MA. A comparison of two tests for equality of two proportions. Am Stat. 1977;31:151-5.
13. Haber M. A comparison of some conditional and unconditional exact tests for 2x2 contingency tables: a comparison of some conditional and unconditional exact tests. Commun Stat Simul Comput. 1987;16(4):999-1013. http://dx.doi.org/10.1080/03610918708812633.
14. Martín Andrés A, Mato AS, Herranz TI. A critical review of asymptotic methods for comparing two proportions by means of independent samples. Commun Stat Simul Comput. 1992;21(2):551-86. http://dx.doi.org/10.1080/03610919208813035.
15. Holmo NF, Ramos GB, Salomao H, et al. Complex segregation analysis of facial melasma in Brazil: evidence for a genetic susceptibility with a dominant pattern of segregation. Arch Dermatol Res. 2018;310(10):827-31. http://dx.doi.org/10.1007/s00403-018-1861-5. PMid:30167816.
16. Tamega AA, Bezerra LVGSP, Pereira FP, Miot HA. Blood groups and discoid lupus erythematosus. An Bras Dermatol. 2009;84(5):477-81. http://dx.doi.org/10.1590/S0365-05962009000500005.
17. Amiri P, Javid AZ, Moradi L, et al. Associations between new and old anthropometric indices with type 2 diabetes mellitus and risk of metabolic complications: a cross-sectional analytical study. J Vasc Bras. 2021;20:e20200236. http://dx.doi.org/10.1590/1677-5449.200236. PMid:34630540.
18. Ludbrook J. Analysis of 2 × 2 tables of frequencies: matching test to experimental design. Int J Epidemiol. 2008;37(6):1430-5. http://dx.doi.org/10.1093/ije/dyn162. PMid:18710887.
19. Oliveira NL, Pereira CAB, Diniz MA, Polpo A. A discussion on significance indices for contingency tables under small sample sizes. PLoS One. 2018;13(8):e0199102. http://dx.doi.org/10.1371/journal.pone.0199102. PMid:30071022.
20. Lloyd CJ. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. Biometrics. 2008;64(3):716-23. http://dx.doi.org/10.1111/j.1541-0420.2007.00936.x. PMid:18047530.
21. Barnard GA. Significance tests for 2 × 2 tables. Biometrika. 1947;34(1-2):123-38. http://dx.doi.org/10.1093/biomet/34.1-2.123. PMid:20287826.
22. Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2 × 2 tables. Stat Med. 2009;28(7):1159-75. http://dx.doi.org/10.1002/sim.3531. PMid:19170020.
23. Goodman LA. On methods for comparing contingency tables. J Roy Stat Soc: Series A (General). 1963;126(1):94-108. http://dx.doi.org/10.2307/2982447.
24. Amiri S, Modarres R. Comparison of tests of contingency tables. J Biopharm Stat. 2017;27(5):784-96. http://dx.doi.org/10.1080/10543406.2016.1269786. PMid:27936354.
25. Ludbrook J. Analysing 2 × 2 contingency tables: which test is best? Clin Exp Pharmacol Physiol. 2013;40(3):177-80. http://dx.doi.org/10.1111/1440-1681.12052. PMid:23294254.
26. Choi L, Blume JD, Dupont WD. Elucidating the foundations of statistical inference with 2 × 2 tables. PLoS One. 2015;10(4):e0121263. http://dx.doi.org/10.1371/journal.pone.0121263. PMid:25849515.
27. Sourial N, Wolfson C, Zhu B, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol. 2010;63(6):638-46. http://dx.doi.org/10.1016/j.jclinepi.2009.08.008. PMid:19896800.
28. Watts DD. Correspondence analysis: a graphical technique for examining categorical data. Nurs Res. 1997;46(4):235-9. http://dx.doi.org/10.1097/00006199-199707000-00009. PMid:9261298.
29. Knapp TR. Treating ordinal scales as ordinal scales. Nurs Res. 1993;42(3):184-6. http://dx.doi.org/10.1097/00006199-199305000-00011. PMid:8506169.
30. Miot HA. Sample size in clinical and experimental studies. J Vasc Bras. 2011;10(4):275-8. http://dx.doi.org/10.1590/S1677-54492011000400001.
31. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74. http://dx.doi.org/10.1177/0962280218784726. PMid:29966490.
32. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ. 1995;311(7013):1145-8. http://dx.doi.org/10.1136/bmj.311.7013.1145. PMid:7580713.
33. Sharpe D. Chi-square test is statistically significant: now what? Pract Assess, Res Eval. 2015;20:8.
34. Carneiro RM, van Bellen B, Santana PRP, Gomes ACP. Prevalence of incidental pulmonary thromboembolism in cancer patients: retrospective analysis at a large center. J Vasc Bras. 2017;16(3):232-8. http://dx.doi.org/10.1590/1677-5449.002117. PMid:29930652.
35. Goodman LA, Kruskal WH. Measures of association for cross classifications. J Am Stat Assoc. 1954;49:732-64.
36. Parshall MB. Unpacking the 2 × 2 table. Heart Lung. 2013;42(3):221-6. http://dx.doi.org/10.1016/j.hrtlng.2013.01.006. PMid:23490241.

37. Miola AC, Miot HA. P-value and effect-size in clinical and experimental studies. J Vasc Bras. 2021;20:e20210038. http://dx.doi.org/10.1590/1677-5449.210038. PMid:34267792.

38. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. Cambridge: Cambridge University Press; 2011. http://dx.doi.org/10.1017/CBO9780511974175.

39. Valenzuela C. 2 solutions for estimating odds ratios with zeros. Rev Med Chil. 1993;121(12):1441-4. PMid:8085071.

40. Lawson R. Small sample confidence intervals for the odds ratio. Commun Stat Simul Comput. 2004;33(4):1095-113. http://dx.doi.org/10.1081/SAC-200040691.

41. Pinto VF. Estudos clínicos de não-inferioridade: fundamentos e controvérsias. J Vasc Bras. 2010;9(3):145-51. http://dx.doi.org/10.1590/S1677-54492010000300009.

42. Mellor K, Eddy S, Peckham N, et al. Progression from external pilot to definitive randomised controlled trial: a methodological review of progression criteria reporting. BMJ Open. 2021;11(6):e048178. http://dx.doi.org/10.1136/bmjopen-2020-048178. PMid:34183348.

43. Willan AR, Thabane L. Bayesian methods for pilot studies. Clin Trials. 2020;17(4):414-9. http://dx.doi.org/10.1177/1740774520914306. PMid:32297539.

44. Thabane L, Lancaster G. A guide to the reporting of protocols of pilot and feasibility trials. Pilot Feasibility Stud. 2019;5(1):37. http://dx.doi.org/10.1186/s40814-019-0423-8. PMid:30858987.

Correspondence
Hélio Amante Miot
Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Faculdade de Ciências Médicas e Biológicas de Botucatu, Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia
Av. Prof. Mário Rubens Guimarães Montenegro, SN, Campus Universitário de Rubião Jr
CEP 18618-000 - Botucatu (SP), Brasil
Tel.: +55 (14) 3811-6015
E-mail: heliomiot@gmail.com

Author information
ACM - Dermatologist; MSc; PhD, Faculdade de Ciências Médicas e Biológicas de Botucatu (UNESP).
HAM – Dermatologist; PhD, Faculdade de Medicina, Universidade de São Paulo (FMUSP); Tenured professor, Faculdade de Ciências Médicas e Biológicas de Botucatu (UNESP).

EDITORIAL

JORNAL **VASCULAR** BRASILEIRO
ISSN 1677-7301 (Online)

Comparação entre variáveis categóricas em estudos clínicos e experimentais

Comparing categorical variables in clinical and experimental studies

Anna Carolina Miola[1], Hélio Amante Miot[1]

Como citar: Miola AC, Miot HA. Comparação entre variáveis categóricas em estudos clínicos e experimentais. J Vasc Bras. 2022;21:e20210225. https://doi.org/10.1590/1677-5449.20210225

Diversos estudos de natureza quantitativa, tanto em ciências biomédicas quanto sociais, utilizam variáveis qualitativas, também chamadas de categóricas, as quais expressam sua grandeza pela frequência em que cada uma de suas categorias ocorre. Variáveis qualitativas são divididas em dicotômicas (por exemplo, sexo, óbito, cura), ordinais (por exemplo, estadiamento neoplásico, amplitude do pulso, classe funcional, fototipo, risco anestésico) ou politômicas/multinominais (por exemplo, orientação sexual, tipagem ABO, estado civil, religião, raça, tipo de aneurisma, tipo de úlcera crônica)[1-3].

Quando se utilizam variáveis qualitativas, o fenômeno mensurado pode ser representado pelo percentual de ocorrência em cada categoria, e sua comparação entre os subgrupos deve ser realizada de acordo com a proporção que cada classe ocupa na amostra[3]. Há extensa literatura a respeito de técnicas de análise estatística de variáveis qualitativas[4-6]; contudo, este texto abordará a comparação de proporções entre variáveis categóricas. A análise comparativa dessas proporções entre subgrupos utiliza conceitos diferentes da chamada estatística paramétrica, apresentando menor poder estatístico (maior erro tipo II) em situações análogas, como quando uma variável quantitativa (por exemplo, idade) é categorizada (por exemplo, < 30 anos, 30–59 anos, ≥ 60 anos)[7,8].

Segundo a estatística frequentista[9], a probabilidade de uma proporção de eventos selecionados aleatoriamente, sem reposição de casos, pode ser generalizada a partir da distribuição qui-quadrado, e o teste qui-quadrado de Pearson baseia-se na diferença entre as frequências encontradas e as idealmente esperadas para cada categoria, podendo ser utilizado para comparar a aderência da amostra a uma distribuição conhecida (por exemplo, comparação com a literatura) ou a independência entre diferentes amostras[10]. Apesar da popularidade do teste qui-quadrado de Pearson, outros métodos como o teste G (razão de verossimilhança) e o teste de Goodman (contrastes de proporções) também são utilizados para comparação de proporções. Todavia, a superioridade absoluta entre eles ainda não foi sistematicamente definida[11-14].

A aderência de uma proporção encontrada pode ser comparada a uma descrição da literatura ou a uma expectativa teórica (por exemplo, expressão de um fenótipo segundo a segregação de um gene)[15]. Exemplificando, Tamega et al.[16] estudaram os tipos sanguíneos ABO e Rh de 69 pacientes com lúpus eritematoso, comparando-os com a frequência esperada dessas categorias entre os doadores de sangue da instituição. O teste qui-quadrado de Pearson (de aderência) resultou p-valor = 0,081 para as tripagens ABO e p-valor = 0,721 para a tipagem Rh, aceitando a hipótese que tais classes de tipos sanguíneos encontrados não divergiam do esperado, na população local[16].

Na pesquisa clínico-epidemiológica, é bastante usual que se apresente uma tabela inicial descritiva com os dados demográficos dos subgrupos, a fim de atestar a sua homogeneidade. Por exemplo, Amiri et al.[17] incluíram 110 casos e 110 controles em um estudo transversal para testar a associação de índices antropométricos e diabetes melito tipo 2. Entre os diabéticos, 75 (51%) eram do sexo feminino, enquanto no grupo controle foram encontradas 72 mulheres (49%). Segundo essa amostra, a diferença de proporção entre os grupos (2%) não foi considerada significativa (p-valor = 0,668) para essa variável dicotômica segundo o teste qui-quadrado de Pearson (de independência).

Apesar de versátil, o teste qui-quadrado de Pearson apresenta performance inadequada (maior erro tipo I)

[1] Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Faculdade de Ciências Médicas e Biológicas de Botucatu, Botucatu, SP, Brasil.
Fonte de financiamento: Nenhuma.
Conflito de interesse: Os autores declararam não haver conflitos de interesse que precisam ser informados.
Submetido em: Dezembro 11, 2021. Aceito em: Janeiro 20, 2022.

O estudo foi realizado Departamento de Dermatologia, Faculdade de Ciências Médicas e Biológicas de Botucatu, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Botucatu, SP, Brasil.

em amostras mais modestas ($n \leq 40$), especialmente nas situações em que > 20% dos valores esperados for ≤ 5, o que é relativamente frequente no cenário da pesquisa biomédica. Diversos procedimentos são recomendados nessa situação, desde a fusão de categorias para aumentar o valor esperado (por exemplo, dicotomizar as cores de pele como branca vs. não branca, agrupar os tipos sanguíneos menos comuns B com AB), ou mesmo o uso de outros testes estatísticos.

Há intensa discussão acadêmica sobre quais estratégias analíticas devem ser utilizadas para as situações em que o teste qui-quadrado de Pearson for contraindicado; da mesma forma, diferentes testes para dados categóricos podem se comportar diversamente, de acordo com a forma que as variáveis são coletadas (aleatórias ou não), já que grande parte dos estudos não possuem estrutura amostral completamente aleatorizada[18-20]. Os testes exatos de Barnard e de Boschloo são exemplos que corrigem essas limitações para tabelas de contingência 2 × 2[21,22]. Já o teste G (com correção de Williams) pode ser utilizado para comparações multinominais em situações de contraindicação do teste qui-quadrado de Pearson[21,23]. Estimativas do p-valor (exato) a partir de reamostragens (*bootstrap*) ou simulação de Monte Carlo também são eficientes para sua estimativa em casos de amostras modestas ou subgrupos com baixa expectativa de ocorrência[19,24].

O teste exato de Fisher é citado por muitos textos como solução para casos em que o teste qui-quadrado de Pearson não seja indicado, porém, ele inflaciona o erro tipo II, além de se basear em um modelo de probabilidade condicional, diferente do que é usualmente proposto em pesquisa biomédica (totais marginais variáveis)[25,26]. Da mesma forma, a correção do teste qui-quadrado de Pearson pelo procedimento de Yates, em tabelas 2 × 2, é excessivamente conservadora. O emprego e as interpretações desses testes devem ser parcimoniosos quando resultarem p-valor próximo ao nível de significância[22,24].

Em desenhos mais complexos, que envolvam a interação de mais de duas variáveis categóricas ou ajuste multivariado cuja variável dependente seja categórica, outros métodos de análise podem ser utilizados, como a regressão de Poisson (*log-linear*), a regressão logística e a regressão multinominal, que, assim como no teste qui-quadrado de Pearson, são penalizadas pela ocorrência de frequências baixas entre os subgrupos. Por outro lado, métodos multivariados, como a análise de correspondência múltipla, não são afetados pelas contingências dos testes de hipóteses e podem substanciar análises exploratórias para dados categóricos[4,27,28]. Em tempo, a problemática ligada à análise de dados ordinais e o cálculo do tamanho amostral para estudos que envolvam proporções já foram abordados anteriormente[2,29-32].

Quando comparações de variáveis multinominais resultam significativas, cabe saber quais das proporções internas apresentam divergência do esperado, tendo em vista que o resultado do teste (por exemplo, o teste qui-quadrado de Pearson) refere-se ao comportamento global das proporções, devendo-se, então, proceder a análise *post hoc* das subcategorias. A análise de resíduos da tabela de contingência (padronizada e ajustada) é uma estratégia muito empregada, que retorna a estatística *Z* (Zres) para cada proporção encontrada, permitindo a comparação múltipla entre elas ao identificar quais variáveis específicas mais contribuem para o resultado encontrado no teste global[33]. A partir da análise de resíduos da Tabela 1, pode-se concluir que pacientes oncológicos oriundos do ambulatório apresentaram mais diagnósticos tomográficos de tromboembolismo pulmonar incidentais que os originários do pronto-socorro, sem diferenças com as proporções encontradas na enfermaria e UTI[34].

Outra opção para a análise dos subgrupos é o teste *lambda* de Goodman e Kruskal que se trata de uma medida de redução proporcional no erro na análise de tabela de contingência para dados multinomiais, indicando até que ponto as categorias e frequências modais para cada valor da variável independente diferem dos valores da variável independente[35]. Da mesma forma, a partição da tabela em subtabelas 2 × 2 pode ser realizada. Contudo, as comparações múltiplas devem ser ajustadas para reduzir o inflacionamento

Tabela 1. Análise de resíduos dos dados de Carneiro et al.[34] quanto à origem dos pacientes oncológicos com tromboembolismo pulmonar (TEP) à tomografia computadorizada de tórax, quando o achado foi incidental ou havia suspeita prévia.

Origem	Sem suspeita de TEP		Com suspeita de TEP	
	(*n* = 48)	Zres (p-valor)	(*n* = 60)	Zres (p-valor)
Ambulatorial	28 (59%)	*+5,1 (<0,001)*	7 (12%)	*-5,1 (<0,001)*
Enfermaria	16 (33%)	-0,2 (0,856)	21 (35%)	+0,2 (0,856)
Unidade de Tratamento Intensivo	2 (4%)	-1,4 (0,161)	7 (12%)	+1,4 (0,161)
Pronto-socorro	2 (4%)	*-4,5 (<0,001)*	25 (43%)	*+4,5 (<0,001)*

p-valor (global) < 0,001; qui-quadrado de Pearson.

do erro tipo I, por exemplo, usando o procedimento de Bonferroni[20].

A pesquisa epidemiológica utiliza frequentemente desfechos dicotômicos (por exemplo, cura, óbito, adoecimento) para a comparação de dois ou mais grupos (por exemplo, placebo vs. tratamento). Devido à característica intrínseca do desenho dos estudos, há crescente tendência que a comparação dessas proporções seja estimada a partir das suas medidas epidemiológicas de efeito, como razão de chances, risco relativo ou razão de prevalências, e não somente pelos testes estatísticos de proporção[36,37]. Tanto o p-valor como o intervalo de confiança para essas associações podem ser calculados diretamente para essas estimativas a partir de modelos de regressão logística, ordinal, multinominal ou de Poisson[38].

A necessidade de ajuste dos resultados por covariáveis de importância no modelo causal (por exemplo, idade, sexo, tabagismo) vem demandando a popularização dessas técnicas de regressão para a análise de dados categóricos, e a contingência diante das amostras modestas ou da raridade de eventos em uma das categorias pode ser transposta por técnicas de *bootstrap*, com mais de 1.000 reamostragens dos dados. Entretanto, como esses métodos ponderam as relações entre as subcategorias, eles não lidam adequadamente quando uma delas é zero, ao contrário das técnicas estatísticas exatas (por exemplo, teste de Barnard).

A Tabela 2 exemplifica formas de análise para comparações de dois tratamentos hipotéticos (cirúrgico vs. convencional) analisados segundo testes de comparação de proporções e modelos de regressão, de acordo com particularidades amostrais. No caso especial, para estimar a dimensão de efeito de um estudo (por exemplo, risco relativo e razão de chances) em que houve zero ocorrências em uma das variáveis categóricas, pode-se recorrer à adição (artificial) de 0,5 unidades nos desfechos de cada grupo[5,39,40].

A comparação de proporções entre grupos também pode ser avaliada de forma uni ou bidirecional (uni/bicaudal), já que muitas avaliações são, por natureza, unidirecionais, como a comparação da taxa de mortalidade em uma doença entre vacinados e não vacinados ou em testes de não inferioridade entre dois tratamentos[41]. Nesses casos, não faz parte da hipótese de pesquisa a possibilidade de que o resultado seja contemplado de forma bidirecional, interessando apenas o efeito em um sentido. Análises unicaudais entre proporções não são consensuais entre os epidemiologistas, porque, apesar de apresentarem maior poder estatístico e demandarem menor amostragem, aumentam a chance de erro tipo I[24]. Análises unicaudais são muito empregadas em estudos de viabilidade (estudos piloto) e em provas de conceito, que ocorrem antes dos ensaios clínicos tradicionais[42-44].

Situações que envolvam dados dependentes devem ser avaliadas pelo teste de McNemar (tabelas 2 × 2), teste Q de Cochran (vários grupos, resposta dicotômica) ou equações de estimativas generalizadas. Tais análises, assim como uso de técnicas de reamostragem, estimativas unicaudais, regressões e análises de variáveis que demandem ajuste multivariado, devem ser supervisionadas por estatístico experiente.

Tabela 2. Exemplos hipotéticos de comparações (bicaudais) da incidência de morte de uma doença tratada com um procedimento cirúrgico ou um tratamento convencional.

Exemplos	Teste estatístico	Estatística/efeito	p-valor
2 mortes em 100 cirurgias (2%)	Qui-quadrado de Pearson	χ^2 = 11,97; GL = 1	<0,001
vs.			
16 mortes em 100 tratamentos convencionais (16%)	Regressão de Poisson (robusta)	RR = 0,13	0,005
		IC 95% = 0,03 a 0,53	
1 mortes em 50 cirurgias (2%)	Barnard	Escore = 2,45	0,016
vs.			
8 mortes em 50 tratamentos convencionais (16%)	Regressão de Poisson	RR = 0,13	0,046
	(robusta; 1000 reamostragens)	IC 95% = 0,01 a 0,43	
Zero mortes em 50 cirurgias (0%)	Barnard	Escore = 2,95	0,004
vs.			
8 mortes em 50 tratamentos convencionais (186%)		RR = 0,06[a]	0,034
		IC 95% = 0,01 a 0,45	

GL = graus de liberdade; RR = risco relativo; IC 95% = intervalo de confiança de 95%. [a]Risco relativo calculado após inclusão de 0,5 unidades no desfecho de cada grupo: 0,5 mortes entre as cirurgias, 8,5 mortes entre os tratamentos convencionais.

Finalmente, a comparação entre variáveis categóricas é uma demanda frequente em estudos biomédicos e que pode resultar em diferentes conclusões inferenciais de acordo com o método analítico empregado, especialmente quando as frequências nos subgrupos forem baixas. A escolha da técnica de análise exige fundamentação teórica, e sua descrição precisa ser justificada na metodologia, quanto aos parâmetros de uso.

REFERÊNCIAS

1. Greenhalgh T. How to read a paper: statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ. 1997;315(7104):364-6. http://dx.doi.org/10.1136/bmj.315.7104.364. PMid:9270463.
2. Miot HA. Analysis of ordinal data in clinical and experimental studies. J Vasc Bras. 2020;19:e20200185. http://dx.doi.org/10.1590/1677-5449.200185. PMid:34211532.
3. Perkins SM. Statistical inference on categorical variables. Methods Mol Biol. 2007;404:73-88. http://dx.doi.org/10.1007/978-1-59745-530-5_5. PMid:18450046.
4. Pereira JCR. Análise de dados qualitativos: estratégias metodológicas para as ciências da saúde humanas e sociais. São Paulo: EdUSP; 1999.
5. Agresti A. An introduction to categorical data analysis. 2nd ed. New Jersey: John Wiley & Sons; 2020.
6. Quinn GP, Keough MJ. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press; 2002. http://dx.doi.org/10.1017/CBO9780511806384.
7. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41. http://dx.doi.org/10.1002/sim.2331. PMid:16217841.
8. Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol. 2011;32(3):437-40. http://dx.doi.org/10.3174/ajnr.A2425. PMid:21330400.
9. Zaslavsky BG. Bayesian versus frequentist hypotheses testing in clinical trials with dichotomous and countable outcomes. J Biopharm Stat. 2010;20(5):985-97. http://dx.doi.org/10.1080/10543401003619023. PMid:20721786.
10. Turner N. Chi-squared test. J Clin Nurs. 2000;9(1):93. PMid:11041649.
11. Goodman LA. On the multivariate analysis of three dichotomous variables. Ajs. 1965;71(3):290-301. http://dx.doi.org/10.1086/224088. PMid:5897475.
12. Eberhardt KR, Fligner MA. A comparison of two tests for equality of two proportions. Am Stat. 1977;31:151-5.
13. Haber M. A comparison of some conditional and unconditional exact tests for 2x2 contingency tables: a comparison of some conditional and unconditional exact tests. Commun Stat Simul Comput. 1987;16(4):999-1013. http://dx.doi.org/10.1080/03610918708812633.
14. Martín Andrés A, Mato AS, Herranz TI. A critical review of asymptotic methods for comparing two proportions by means of independent samples. Commun Stat Simul Comput. 1992;21(2):551-86. http://dx.doi.org/10.1080/03610919208813035.
15. Holmo NF, Ramos GB, Salomao H, et al. Complex segregation analysis of facial melasma in Brazil: evidence for a genetic susceptibility with a dominant pattern of segregation. Arch Dermatol Res. 2018;310(10):827-31. http://dx.doi.org/10.1007/s00403-018-1861-5. PMid:30167816.
16. Tamega AA, Bezerra LVGSP, Pereira FP, Miot HA. Blood groups and discoid lupus erythematosus. An Bras Dermatol. 2009;84(5):477-81. http://dx.doi.org/10.1590/S0365-05962009000500005.
17. Amiri P, Javid AZ, Moradi L, et al. Associations between new and old anthropometric indices with type 2 diabetes mellitus and risk of metabolic complications: a cross-sectional analytical study. J Vasc Bras. 2021;20:e20200236. http://dx.doi.org/10.1590/1677-5449.200236. PMid:34630540.
18. Ludbrook J. Analysis of 2 × 2 tables of frequencies: matching test to experimental design. Int J Epidemiol. 2008;37(6):1430-5. http://dx.doi.org/10.1093/ije/dyn162. PMid:18710887.
19. Oliveira NL, Pereira CAB, Diniz MA, Polpo A. A discussion on significance indices for contingency tables under small sample sizes. PLoS One. 2018;13(8):e0199102. http://dx.doi.org/10.1371/journal.pone.0199102. PMid:30071022.
20. Lloyd CJ. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. Biometrics. 2008;64(3):716-23. http://dx.doi.org/10.1111/j.1541-0420.2007.00936.x. PMid:18047530.
21. Barnard GA. Significance tests for 2 × 2 tables. Biometrika. 1947;34(1-2):123-38. http://dx.doi.org/10.1093/biomet/34.1-2.123. PMid:20287826.
22. Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2 × 2 tables. Stat Med. 2009;28(7):1159-75. http://dx.doi.org/10.1002/sim.3531. PMid:19170020.
23. Goodman LA. On methods for comparing contingency tables. J Roy Stat Soc: Series A (General). 1963;126(1):94-108. http://dx.doi.org/10.2307/2982447.
24. Amiri S, Modarres R. Comparison of tests of contingency tables. J Biopharm Stat. 2017;27(5):784-96. http://dx.doi.org/10.1080/10543406.2016.1269786. PMid:27936354.
25. Ludbrook J. Analysing 2 × 2 contingency tables: which test is best? Clin Exp Pharmacol Physiol. 2013;40(3):177-80. http://dx.doi.org/10.1111/1440-1681.12052. PMid:23294254.
26. Choi L, Blume JD, Dupont WD. Elucidating the foundations of statistical inference with 2 × 2 tables. PLoS One. 2015;10(4):e0121263. http://dx.doi.org/10.1371/journal.pone.0121263. PMid:25849515.
27. Sourial N, Wolfson C, Zhu B, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol. 2010;63(6):638-46. http://dx.doi.org/10.1016/j.jclinepi.2009.08.008. PMid:19896800.
28. Watts DD. Correspondence analysis: a graphical technique for examining categorical data. Nurs Res. 1997;46(4):235-9. http://dx.doi.org/10.1097/00006199-199707000-00009. PMid:9261298.
29. Knapp TR. Treating ordinal scales as ordinal scales. Nurs Res. 1993;42(3):184-6. http://dx.doi.org/10.1097/00006199-199305000-00011. PMid:8506169.
30. Miot HA. Sample size in clinical and experimental studies. J Vasc Bras. 2011;10(4):275-8. http://dx.doi.org/10.1590/S1677-54492011000400001.
31. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74. http://dx.doi.org/10.1177/0962280218784726. PMid:29966490.
32. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ. 1995;311(7013):1145-8. http://dx.doi.org/10.1136/bmj.311.7013.1145. PMid:7580713.

33. Sharpe D. Chi-square test is statistically significant: now what? Pract Assess, Res Eval. 2015;20:8.

34. Carneiro RM, van Bellen B, Santana PRP, Gomes ACP. Prevalence of incidental pulmonary thromboembolism in cancer patients: retrospective analysis at a large center. J Vasc Bras. 2017;16(3):232-8. http://dx.doi.org/10.1590/1677-5449.002117. PMid:29930652.

35. Goodman LA, Kruskal WH. Measures of association for cross classifications. J Am Stat Assoc. 1954;49:732-64.

36. Parshall MB. Unpacking the 2 × 2 table. Heart Lung. 2013;42(3):221-6. http://dx.doi.org/10.1016/j.hrtlng.2013.01.006. PMid:23490241.

37. Miola AC, Miot HA. P-value and effect-size in clinical and experimental studies. J Vasc Bras. 2021;20:e20210038. http://dx.doi.org/10.1590/1677-5449.210038. PMid:34267792.

38. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. Cambridge: Cambridge University Press; 2011. http://dx.doi.org/10.1017/CBO9780511974175.

39. Valenzuela C. 2 solutions for estimating odds ratios with zeros. Rev Med Chil. 1993;121(12):1441-4. PMid:8085071.

40. Lawson R. Small sample confidence intervals for the odds ratio. Commun Stat Simul Comput. 2004;33(4):1095-113. http://dx.doi.org/10.1081/SAC-200040691.

41. Pinto VF. Estudos clínicos de não-inferioridade: fundamentos e controvérsias. J Vasc Bras. 2010;9(3):145-51. http://dx.doi.org/10.1590/S1677-54492010000300009.

42. Mellor K, Eddy S, Peckham N, et al. Progression from external pilot to definitive randomised controlled trial: a methodological review of progression criteria reporting. BMJ Open. 2021;11(6):e048178. http://dx.doi.org/10.1136/bmjopen-2020-048178. PMid:34183348.

43. Willan AR, Thabane L. Bayesian methods for pilot studies. Clin Trials. 2020;17(4):414-9. http://dx.doi.org/10.1177/1740774520914306. PMid:32297539.

44. Thabane L, Lancaster G. A guide to the reporting of protocols of pilot and feasibility trials. Pilot Feasibility Stud. 2019;5(1):37. http://dx.doi.org/10.1186/s40814-019-0423-8. PMid:30858987.

Correspondência
Hélio Amante Miot
Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Faculdade de Ciências Médicas e Biológicas de Botucatu, Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia
Av. Prof. Mário Rubens Guimarães Montenegro, SN, Campus Universitário de Rubião Jr
CEP 18618-000 - Botucatu (SP), Brasil
Tel.: (14) 3811-6015
E-mail: heliomiot@gmail.com

Informações sobre os autores
ACM - Dermatologista; Mestre; PhD, Faculdade de Ciências Médicas e Biológicas de Botucatu (UNESP).
HAM - Dermatologista; PhD, Faculdade de Medicina, Universidade de São Paulo (FMUSP); Livre-docente, Faculdade de Ciências Médicas e Biológicas de Botucatu (UNESP).