Test of association: which one is the most appropriate for my study?*

David Alejandro Gonzalez-Chica[1]
Rodrigo Pereira Duquia[2]
Jeovany Martínez-Mesa[3]
João Luiz Bastos[1]
Renan Rangel Bonamigo[2]

DOI: http://dx.doi.org/10.1590/abd1806-4841.20154289

Abstract: BACKGROUND: Hypothesis tests are statistical tools widely used for assessing whether or not there is an association between two or more variables. These tests provide a probability of the type 1 error (p-value), which is used to accept or reject the null study hypothesis.
OBJECTIVE: To provide a practical guide to help researchers carefully select the most appropriate procedure to answer the research question. We discuss the logic of hypothesis testing and present the prerequisites of each procedure based on practical examples.
Keywords: Data analysis; Association; Epidemiology and biostatistics; Hypothesis testing; Statistical methods and procedures

INTRODUCTION

As shown in previous publications, every scientific study should begin with a clearly defined research question. [1-3] In this paper, we will cover the basic assumptions of statistical analysis that are followed in bivariate association tests (which involve one exposure and one outcome) and review the general principles underlying their implementation. We intend to provide readers with a quick and simple guide that will help them choose the most appropriate tests for each situation or research question.

WHAT ARE HYPOTHESIS TESTS?

Hypothesis is defined as a "statement that can be questioned or tested, and that may be refuted in scientific studies."[2] Along with the null hypothesis (H_0 - the original assumption of no difference or no association that is accepted as being true for a given situation), there is the alternative hypothesis (H_A - an additional explanation for the same situation, which may replace H_0 and needs to be tested). For example, in the randomized clinical trial (RCT) by Bagatin et al., H_0 states that both oral isotretinoin (ISO) and topical retinoic acid 0.05% (RA- 0.05%) have the same effect on several outcomes related to photoaging. [4] By contrast, H_A tested in the aforementioned study assumes that the effect of isotretinoin is better than the effect of topical RA on photoaging.

Usually, when working with hypothesis testing, what the investigator needs to know is whether a particular outcome (e.g., the size of an injury, blood marker levels, etc.) is different when the intervention group and the control group are compared (in experimental studies) or when exposed and unexposed subjects are contrasted (in observational studies).

When faced with the decision of rejecting the H_0, researchers need to define a priori the maximum acceptable probability of type I or "alpha" error (probability of rejecting H_0 based on the sample results, when H_0 is actually true in the target population) that will be considered as acceptable. Usually, type I error for two-tailed tests is set at 5% (and at 2.5% for one-tailed tests). In scientific papers, this probability is called "p-value" and is used to determine whether a result is "statistically significant" or not.

Received on 07.12.2014.
Approved by the Advisory Board and accepted for publication on 09.03.2015.
* Study conducted at the Federal University of Santa Catarina (UFSC) – Florianópolis (SC), Brazil.

Financial Support: None.
Conflict of Interest: None.

[1] Federal University of Santa Catarina (UFSC) – Florianópolis (SC), Brazil.
[2] Federal University of Health Sciences, Porto Alegre (UFCSPA) – Porto Alegre (RS), Brazil.
[3] Latin American Cooperative Oncology Group (LACOG) – Porto Alegre (RS), Brazil.

Statistical tests are nothing but tools that help researchers to find this p-value, based on a statistical formula and a reference table of probabilities, which correspond to the type I error. These formulas and tables can be easily found in statistics textbooks.[5,6] Figure 1 shows the sequence of steps involved in selecting the test and the estimating of p-values.

To estimate the p-value, the results of the statistical test need to be combined with the number of the degrees of freedom. These, in turn, result from a combination between the number of individuals evaluated and/or the number of groups/categories being compared. The degrees of freedom are closely related to the type 1 error, because the smaller the number of individuals and/or the greater the number of groups being compared, the smaller is the probability of the result being "statistically significant" (p-value < 0.05 or <5%).

IMPORTANT ASPECTS WHEN SELECTING THE STATISTICAL TEST

There are several aspects to which the researcher needs to be aware when selecting the statistical test to be used. The first aspect relates to the type of data at hand, which may be "independent" or "paired".

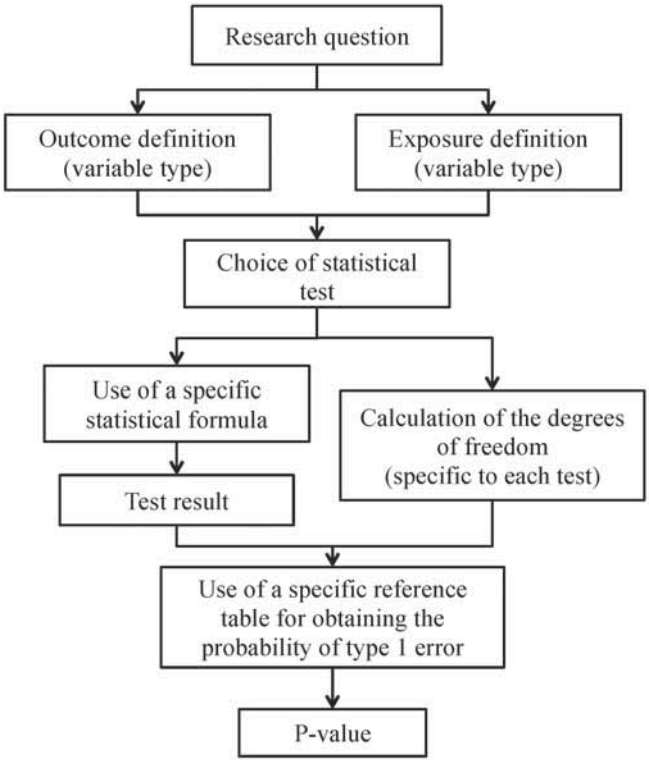

Figure 1: Sequence of steps involved in the estimation of the p-value in statistical analysis

Data are considered independent when the parameters found in an individual do not depend upon the values observed for another respondents in the sample. For example, in the RCT by Bagatin et al[4], the authors assume that each individual shows a certain response to photoaging, and, in theory, this response is not influenced by the response shown by other participants in the study. The same applies to cross-sectional and cohort studies, as well as to unmatched case-control investigations. This is the case of the cross-sectional study by Duquia et al. [7], which assessed only one person per household. Selecting more than one individual per household may affect the independence of observations, given that co-residents may present similar characteristics because individuals influence each other - for example, if a woman wears sunscreen, it is more likely that her spouse and/or children wear it as well.

However, when the results or the selection of an individual are related to the results or the selection of one or more participants in the same study, the data is considered to be "paired", as in the case of matched case-control studies.[8,9] The same principle applies to before-and-after studies, such as the study conducted in Spain to assess the effects of an educational intervention on the adoption of practices to prevent skin cancer in children. [10] In this type of study, all individuals receive the same intervention (there is no other comparison group), and the results observed "after" the intervention was implemented are compared with the information obtained in the baseline, that is, the effects of the intervention are dependent upon the individuals' previous conditions.

The next step in selecting the statistical test considers the type of variable through which the outcome and the exposure variable were measured. At this stage, several requirements for the choice of the statistical test will have to be considered.

STATISTICAL TESTS FOR NUMERIC OUTCOMES

The main criterion used to determine the type of test that should be selected for analyzing numeric outcomes is the symmetry of the variable. In this case, "parametric" tests are the most appropriate ones. A variable is considered to be symmetrical (or "normal") when its mean (or average) and median are similar, and the dispersion of values (distribution of data) is the same on the left and on the right of these measures of central tendency (68% of observations are within ± 1 standard deviation and 95% are within ± 2 standard deviations).[5,6] In the case of asymmetrical outcomes, the researcher can use "nonparametric" tests or try to transform them into symmetrical outcomes (by using the natural logarithm, for example).

Figure 2 shows parametric and nonparametric

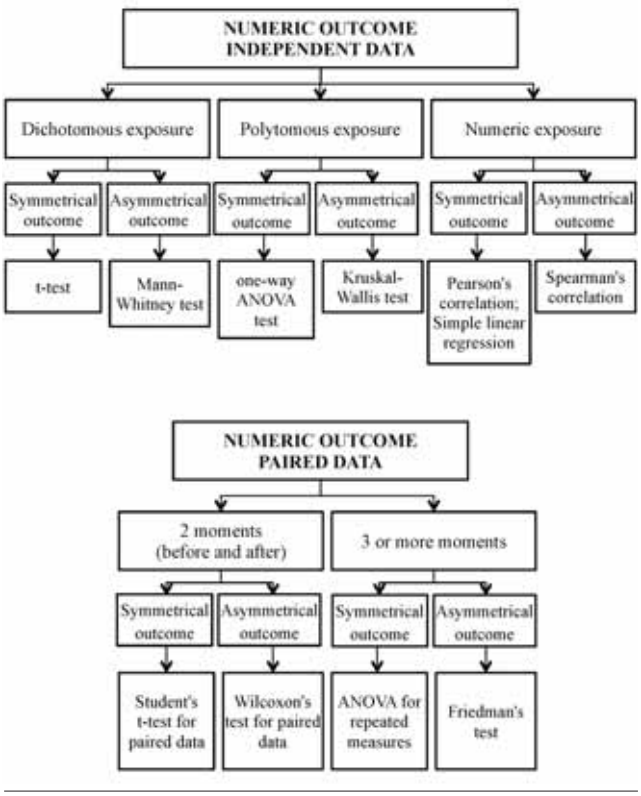

Figure 2: Flowchart for selecting a statistical test for numerical outcomes

test options for numeric outcomes, both for the analysis of independent and paired data.

Both the t-test and the ANOVA test have an additional prerequisite: the variance (or standard deviation) of the outcome should be homogeneous between the groups being compared. There are specific tests that can be used to check this assumption, such as the Bartlett's test (p-value <0.05 indicates heterogeneity of variances between groups). If the variances are not homogeneous, the use of a nonparametric test is recommended (even if the outcome is symmetric), as the p-value resulting from the t-test or the ANOVA test may be biased. However, the t-test is considered to be "robust", because if the number of subjects in a sample is greater than 100 (evenly distributed into the exposed and unexposed groups) and the outcome is symmetric, the p-value will be reliable, even if the variance is heterogeneous between groups.

In both cases the H_A tested is that the mean outcome is different between the categories of the exposure variable. In the case of polytomous exposures, the H_A of the ANOVA and the Kruskal-Wallis tests is that there is a significant difference between the mean outcome of at least two exposure categories (heterogeneity test). For example, in the RCT by Bagatin et al.,[4] the authors could have compared the effects of ISO and two RA concentrations (0.05% and 0.025%) on type I collagen density. A p-value <0.05 in the ANOVA test would confirm the H_A of the study. However, the resulting p-value would not indicate which of the interventions is the best one. Although the inspection of the mean outcome in each category would help to determine these differences, a statistical confirmation using *post-hoc* tests (such as the Bonferroni's, Scheffé's, Newman-Keuls or Duncan's test) is recommended. Post-hoc tests would compare the mean type I collagen in all possible combinations (ISO vs. RA-0.05%; ISO vs RA-0.025%, and RA-0.05% vs. RA-0.025%), in order to identify which groups were statistically different from each other. Given that *post-hoc* tests include a correction for the number of comparisons made (in addition to taking into account the number of individuals in each category), they are more conservative (less likely to show a significant p-value, even if the result of the ANOVA test was <0.05).

When the independent variable is ordinal and the H_A postulates the existence of a trend regarding the outcome (increasing or decreasing the mean outcome according to the categories of the exposure variable), the researcher might decide to use a test of linear trend, instead of a heterogeneity test.

As illustrated in figure 2, when the exposure and the outcome variables are numeric, the researcher can choose to use Pearson's correlation test and/or simple linear regression. In these cases, the exposure variable must also be symmetrical. Spearman's correlation should be used when the exposure and the outcome variables are not symmetrical, or when they are symmetrical but there is not a linear relationship between the variables.

Figure 3 clarifies what is estimated by Pearson's correlation and linear regression. Each subject has an exposure value (X-axis = waist circumference) and an outcome value (Y-axis = body mass index) which can be plotted as a set of point or observations (scatter plot). Each point in this graph is called an "observed" value of the subject. Based on the set of observed values, it is possible to estimate a "prediction line": the expected outcome values for each value of the exposure variable. The difference between the "observed" values and the "predicted" values is called "residual" value. Given that the prediction line crosses the set of points in the scatter plot, the residuals will always have positive and negative values. These values should be normally distributed above, below and along this straight line (which is termed homoscedasticity). Both Pearson's correlation and simple linear regression tests base their estimates on this information, and the prerequisites for both tests include:

1. An approximately linear relationship between

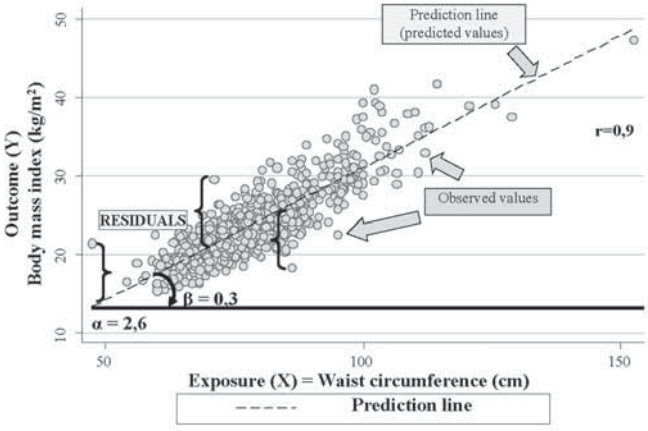

Figure 3: Scatter diagram to demonstrate the association between waist circumference (X-axis) and body mass index (Y-axis). r = correlation coefficient; a = intercept; ß = linear regression coefficient

exposure and outcome: the prediction line should be straight and have a non-zero slope (the steepness does not matter);

2. A symmetrical distribution of theresidual values; and

3. A symmetrical distribution of the outcome and a homogeneous distribution of the residuals along the exposure values (homoscedasticity).

Based on the same data of Figure 3, it is possible to estimate the parameters of the correlation (r = correlation coefficient) and linear regression (α = intercept; β = linear regression coefficient). When testing the association between two numeric variables, the H_A is that both "r" and "β" are different from zero. Chart 1 explains the meaning of these parameters.

Figure 4 shows some instances in which the aforementioned prerequisites are not fulfilled: use of asymmetric variables (exposure and/or outcome) and absence of a linear relationship (despite the existence of symmetric variables). In the three examples provided, it is possible to mathematically estimate a prediction line and estimate "r", "α" and "β", as well as the 95% confidence intervals and corresponding p-values. However, because the prerequisites are not fulfilled, these parameters may be biased.

In the case of paired data, the H_A of the paired t-test is that the difference in means before (baseline or "T0") and after the intervention (end of study or "T1") is different from zero (the same principle applies to

Chart 1: Interpretation of the parameters evaluated to analyze the association between two numeric variables

Parameter	Interpretation	What is it for?	Possible values
Correlation coefficient (r)*	Measurement of linear relationship between two numeric variables, which can be positive (as one variable increases the other variable also increases) or negative (as one variable increases the other variable decreases).	It measures the extent to which the "observed" values approximate the prediction line. It provides information on the direction of association between the variables	From -1.0 (perfect negative correlation) to +1.0 (perfect positive correlation) r = 0 indicates no linear relationship between the two variables. Intermediate values can be ranked as strong (r = 0.7-0.9), moderate (r = 0.4-0.6) or weak correlation (r = 0.1-0.3).
Intercept (α)	Outcome value when the exposure value is zero	It serves to fit an imaginary horizontal line, which is necessary to estimate β	It depends on the parameters of the variables and can fluctuate from -∞ to + ∞
Linear regression coefficient (β)	It represents how much the outcome changes (increase or decrease) with each one-unit increment in the exposure variable.	It determines the prediction line slope in relation to the horizontal line fit by α. It provides information on the direction of association between the variables, as well as on the strength (intensity) of this relationship	It depends on the parameters of the variables and can fluctuate from -∞ to + ∞.

** "r" values should not be interpreted as "strength" of association, given that different slopes in the prediction line (different "β" values, indicating different strength of association) may have the same "r" value*

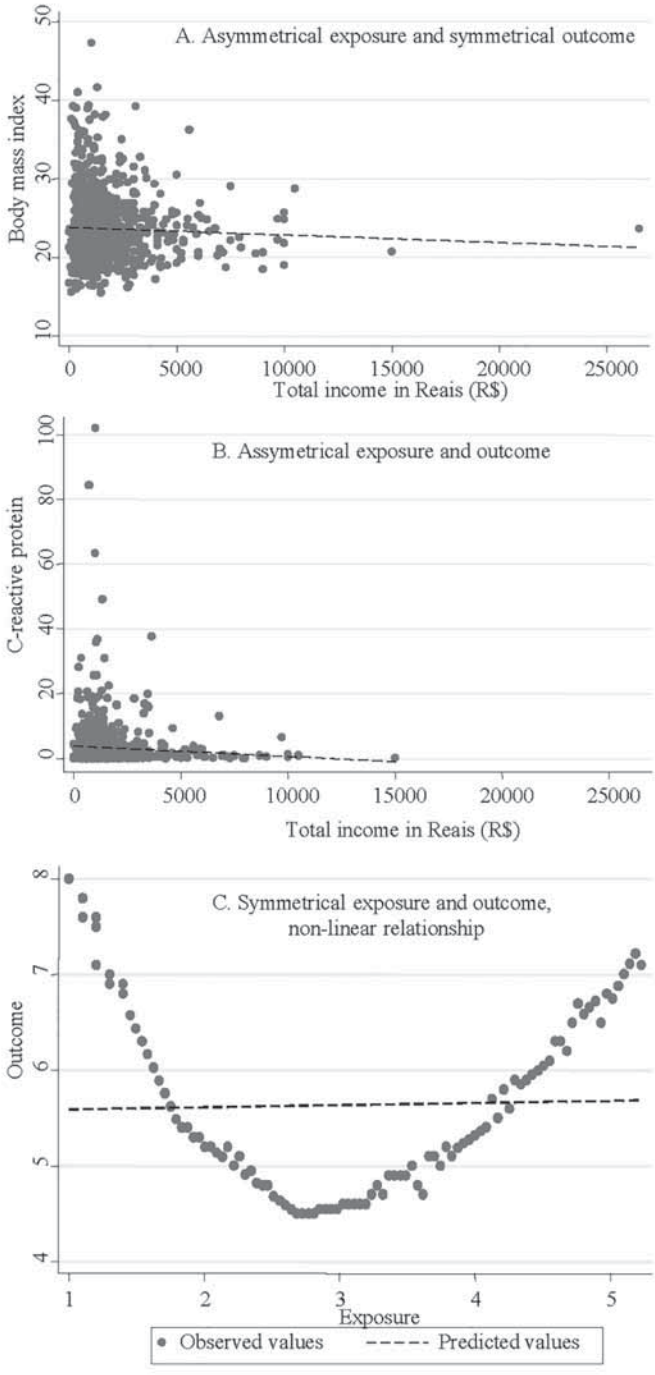

Figure 4: Scatter diagram to illustrate the association between two numeric variables that do not fulfill the prerequisites for simple linear regression or Pearson's correlation

matched case-control studies). In the case of more than one assessment after intervention (T1, T2, T3, etc.), the H_A is that there is a difference between at least two moments (not necessarily in relation to T0). The same requirements of outcome symmetry and homogeneity of variances apply for paired data to allow the use of parametric tests.

STATISTICAL TESTS FOR CATEGORICAL OUTCOMES

The Pearson's chi-square test is used in the case of categorical outcomes, regardless of the number of categories of the outcome or the exposure variables. For instance, when evaluating the association between gender and use of sunscreen, the chi-square test contrasts the "observed" numerical values (absolute frequencies) that were arranged in a contingency table (table containing the number of: sick and exposed subjects; sick and unexposed subjects; healthy, exposed subjects; and healthy, unexposed subjects) and the "expected" values, which correspond to the frequency distribution reflecting no association between the variables (if H_0 is true) (Table 1).[7]

The fundamental requirement of the Pearson's chi-square test is that no expected value is equal to 0. If this occurs, a redefinition of the research question may be necessary and/or the researcher can also consider regrouping the categories of the outcome and/or exposure variables. The second basic requirement of the chi-square test is that the frequencies in the contingency table may not be lower than five in more than 20% of cases (none of the expected values for dichotomous exposure and outcome measurements

Table 1: Contingency tables for testing the association between gender and use of sunscreen at the beach (adapted from Duquia et al. J Am Acad Dermatol.

	A. Observed values Use sunscreen at the beach (Outcome)		
Gender (Exposure)	**No**	**Yes**	**Total**
Male	255	195	450
Female	102	359	461
Total	357	554	911
	B. Expected Values Use sunscreen at the beach (Outcome)		
Gender (Exposure)	**No**	**Yes**	**Total**
Male	176	274	450
Female	181	280	461
Total	357	554	911

Test result = 114; Degrees of freedom = 1; P-value <0.001

Adapted from: Duquia RP, 2007. [7]

may be <5). When this occurs, the chi-square test with Yates' continuity correction (provided that the total sample size is greater than 20) or Fisher's exact test should be used.

The H_A in the aforementioned cases is that the observed frequency of the outcome is different between at least two categories of the exposure variable (chi-square heterogeneity test). When the outcome is dichotomous, and exposure is an ordinal variable, the chi-square test for trend might be used.

For paired data, if both the exposure and outcome were dichotomous, the McNemar's chi-square test should be used. In the case of polytomous variables, tests to assess marginal homogeneity (Stuart-Maxwell or Bhapkar) are the recommended ones.

COMBINING STATISTICAL TESTS

Frequently, the use of more than one statistical test may be necessary in a single research project, due to the different hypotheses being tested. This is the case of the RCT conducted by Muller et al. [11], which investigated the effects of an intervention to improve patients' knowledge of skin lesions suspected of being melanoma. Different tests may also be combined to analyze independent and paired data, as can be seen in the study by Bagatin et al.[4] In such cases, a thorough consideration of which tests should be used to test not only the primary association of interest, but also secondary associations, is essential. When secondary associations are tested, they must also be based on sound theory and should be presented as such by the study authors.

Finally, both observational and intervention studies may require the use of more complex analysis procedures to assess the existence of confounding factors or interaction.[8,9] In addition to selecting the appropriate statistical test, different assumptions must be checked to avoid the estimation of biased type 1 error values, which not only affects the internal validity of the study, but also the extrapolation of the results to the reference population. ❑

REFERENCES

1. Bastos JL, Duquia RP, González-Chica DA, Mesa JM, Bonamigo RR. Field work I: selecting the instrument for data collection. An Bras Dermatol. 2014;89:918-23.
2. Martínez-Mesa J, González-Chica DA, Bastos JL, Bonamigo RR, Duquia RP. Sample size: how many participants do I need in my research? An Bras Dermatol. 2014;89:609-15.
3. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014;89:280-5.
4. Bagatin E, Guadanhim LR, Enokihara MM, Sanudo A, Talarico S, Miot HA, et al. Low-dose oral isotretinoin versus topical retinoic acid for photoaging: a randomized, comparative study. Int J Dermatol. 2014;53:114-22.
5. Altman DG, Machin D, Bryan TN, Gardner MJ. Statistics with confidence: confidence intervals and statistical guidelines. London, UK: BMJ Publishing Group; 2000.
6. Bland M. An introduction to medical statistics. 3rd ed. Oxford, UK: Oxford University Press; 2000.
7. Duquia RP, Baptista Menezes AM, Reichert FF, de Almeida HL Jr. Prevalence and associated factors with sunscreen use in Southern Brazil: A population-based study. J Am Acad Dermatol. 2007;57:73-80.
8. Fletcher RH, Fletcher SW, Wagner EH. Epidemiologia Clínica: Elementos Essenciais. 3.ed. Porto Alegre: Artmed; 2003.
9. Rothman K. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
10. Gilaberte Y, Alonso JP, Teruel MP, Granizo C, Gállego J. Evaluation of a health promotion intervention for skin cancer prevention in Spain: the SolSano program. Health Promot Int. 2008;23:209-19.
11. Müller KR, Bonamigo RR, Crestani TA, Chiaradia G, Rey MC. Evaluation of patients' learning about the ABCD rule: A randomized study in southern Brazil. An Bras Dermatol. 2009;84:593-8.

MAILING ADDRESS:
David Alejandro Gonzalez-Chica
Departamento de Nutrição
Centro de Ciências da Saúde
Universidade Federal de Santa Catarina
Trindade
88040-970 - Florianópolis - SC
Brazil
E-mail: david.epidemio@gmail.com

How to cite this article: Gonzalez-Chica DA, Bastos JL, Duquia RP, Bonamigo RR, Martínez-Mesa J. Tests of association: which one is the most appropriate for my study? An Bras Dermatol. 2015;90(4):523-8.