Levene's Test for Verifying Homoscedasticity Between Groups in Quasi-Experiments in Social Sciences

César Augusto Cardeña Ojeda[1]

[1] *Escuela Normal de Ticul, Ticul, Mexico. cesarcardenaojeda@gmail.com*

KEYWORDS
"Homoscedasticity", "Internal validity", "Levene test", "Quasi-experiment", "Control Group"

ABSTRACT
One of the most used quantitative research designs in the social sciences field is the quasi-experiment with an experimental group and a control group. Its widespread use can be explained, on one hand, by the fact that this design offers greater internal control than pre-experimental designs, which lack a control group. On the other hand, it poses fewer methodological challenges compared to true experiments, whose operational demands often exceed or do not align with the natural behavior of groups in educational populations. Given its prevalence in research studies, it is important to highlight that one of the weaknesses of this design lies precisely in the determination of the treatment and control groups. These groups are often selected based on the researcher's convenience or even inferred through common sense, that is, without a scientifically valid criterion. To address this issue, this article presents Levene's test as an option for verifying homoscedasticity between these two groups, which can contribute to methodological rigor in this aspect that substantially impacts the internal validity of the research findings.

1. Introduction

One of the most widely used designs in quantitative research, particularly in the social and behavioral sciences, is the quasi-experimental design, which includes an experimental group and a control group [1].

The predominance of quasi-experimental studies over true experimental ones can be explained by contrasting certain conditions inherent in both designs with the nature of social subjects, that is, individuals and human groups.

For example, a true experiment requires at least three essential conditions to be met: a) the ability to manipulate an independent variable, b) the ability to randomly assign individuals to study groups, and c) the ability to randomly assign which group or groups will serve as the experimental and control groups [1].

Furthermore, the Solomon design is often used in true experiments, which requires four analysis groups with populations equivalent in their expression of the study variable [2] [3].

This can be a challenge for many educational researchers, who often lack the ability to decide on the number or composition of the groups they work with.

Table I. "Solomon" Research Design

	Pretest	Treatment	Postest
Group 1	O1	X	O2
Group 2	O3		O4
Group 3		X	O5
Group 4			O6

Additionally, it may be necessary to match the size of the groups that make up the study, a process that Mohammad Namakforoosh calls "Matching" [4] either by choosing identical subjects for each group (precise method), or from of the treatment of the mean or variance of an entire group (frequency distribution method) as done by Regression Discontinuity Design and Propensity Score Matching [5].

Once again, this is outside the possibilities of most educational researchers, as they cannot decide on the subjects to whom they have access in their ordinary work.

Finally, it must be considered that a true experiment should proceed without anticipating any disturbance or

atypical reaction in the behavior of the study subjects that could significantly affect the internal validity of the data collected from them. This is to say that, while measurement error is technically unavoidable in social and psychological studies [6][7], the experiment should be conducted under an action plan that objectively anticipates conditions to reduce it to the smallest possible extent [8].

From this, we can recognize that fulfilling the characteristics of a true experiment is complicated in social and behavioral studies, mainly because it is often impossible to subject the process of determining individuals or groups for treatment and control to randomness or manipulation without disturbing their homeostasis or equilibrium.

On the other hand, a quasi-experiment lacks one or more of the three characteristics or demands mentioned earlier [9]. For instance, it may be conducted with two groups, one experimental and one control, and decisions about selecting subjects or groups for study can be made without using random techniques—for example, when subjects are chosen based on convenience.

It may also be necessary to resort to quasi-experimental designs in cases where there are ethical, political, or operational restrictions on selecting subjects or forming groups, even if random selection would not be expected to cause significant issues [5].

Considering this, the predominance of quasi-experimental designs over true experiments can be explained briefly. It is important to note that researchers who recognize that the conditions of their study problem call for a quasi-experiment instead of a true experimental design should not be discouraged by the fact that quasi-experiments provide less control over internal validity than true experiments [10].

The scientific research method is driven by logical reasoning and rationale—the foundation of actions—rather than by satisfying expectations regarding the scope of the study.

Therefore, when a quasi-experiment is determined to be the appropriate method, it should be conducted with the same rigor as any scientific design, and the researcher should dedicate full attention and effort to providing as much foresight and support as possible for the research method.

However, this is not always the case. For example, a review of 62 research theses written between 2011 and 2023 at three higher education institutions, in which a quasi-experiment was conducted, revealed that only four of them (6%) reported taking any precautions when defining the two study groups. Of those four, two used a formal technique for group determination, while the other two merely stated that their groups were equivalent without explaining how this conclusion was reached.

Given this context, the issue of determining the control group in a quasi-experiment was addressed, as this is often done without certainty about the equivalence of the groups being compared.

The determination of groups is often guided by common sense or ordinary reasoning [11] which, while well-intentioned, deviates from the principles of scientific activity [12].

For instance, a researcher teaching Subject X to Group A in a specific grade level at a particular institution may assume that Group B, which is in the same grade level at the same institution and being taught Subject X by the same teacher, constitutes an adequate control group for conducting a quasi-experiment on a psychopedagogical topic. This assumption, which seems quite logical from a common-sense perspective, is because both groups (A and B) share many important contextual characteristics: the same instructor for the same subject (the researcher), the same school environment, a relatively narrow age range among the students, and similar socioeconomic conditions that likely led them to attend the same school.

However, this assumption of equality overlooks potential factors that could invalidate the conclusions drawn from it, such as Intercurrent Variables, which are often unknown and not directly observable, or Extraneous Variables, which may not have been anticipated in the study, often due to insufficient review of the theoretical framework or methodological design [13].

Some examples of these invalidating factors or variables in the hypothetical case proposed would be: the individual inclination or liking (or disliking) of the students towards the subject taught by the teacher-researcher, the quality of the socio-emotional integration (or disintegration) in each group, or the possibility that such groups have been determined by the school authority based on the academic performance of its members, concentrating the students with the best performances separately from the less outstanding ones. All these

possibilities can determine heterogeneous behaviors between both groups, despite their apparent coincidences.

As is evident, supporting the determination of the control group on logical but imprecise or erroneous reasoning can lead to measurement biases that lead to errors in hypothesis I (A) or II (B), that is, rejecting the null hypothesis when it is true, or retaining it. the null hypothesis when it is false, respectively, which is equivalent to detracting from the validity of the results and conclusions obtained in the research work.

On the other hand, taking due precautions to determine an adequate control group allows us to have data that increases the level of inferential analysis on the results that must be attributed to the experiment or quasi-experiment; In other words: this “increases the possibilities of obtaining internal validity in the study” [1].

That is why this text aims to propose a technical resource to corroborate that two groups to be confronted are in equal conditions to be -precisely- compared for study purposes.

2. Development

The tests commonly used for this purpose are based on analyzing the assumption of equal variances between the groups to be compared, also known as "Homoscedasticity" [14].

This can be formulated as a variance ratio equal to one, as expressed in the following hypothesis matrix:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

Figure 1. Null and Alternative hypothesis for the homoscedasticity assumption

These tests, based on the median of the variance, are more used than those based on the arithmetic mean, in part because they offer more precise data to estimate the standard error [15] [16].

Among the most practical tools for examining the homoscedasticity assumption are the F-test, Bartlett's test, and Levene's test.

The F-test (or variance ratio test) is a very powerful option for analyzing homoscedasticity; however, it is highly sensitive to deviations from the normality of population distributions. Therefore, it is not recommended when there is uncertainty about the distribution of the data or when the scores may be affected by arithmetic regression, which can alter the normality condition from one measurement to another.

Bartlett's test is useful when comparing two or more groups, and it has the advantage of not requiring the groups to be of equal size, thus eliminating the need to apply matching techniques. However, like the F-test, it is substantially affected by uncertainty regarding the normality of the population.

On the other hand, Levene's test, in addition to being able to compare two or more groups, is not severely impacted by the uncertainty of the normality assumption, making it a versatile and reliable test [17].

In this regard, Conover, Johnson and Johnson [16] conducted simulated exercises on 56 different homoscedasticity tests, intentionally incorporating violations of the normality assumption: they found that Levene's test demonstrated the greatest power and stability in maintaining the nominal rate of Type I error.

Nonetheless, it is important to clarify that experts do not consider Levene's test to be the immediate or infallible option for verifying homoscedasticity.

Despite its robustness in the face of population variability, it remains important to explore tests for assessing the normality assumption as well.

Calculating Levene's Test Using SPSS

The Statistical Package for Social Sciences (SPSS) has become a very popular tool for performing statistical

calculations in various areas of social sciences research. Given its widespread use, the following outlines the general process for performing Levene's homoscedasticity test using this software:

Step 1. Set the level of statistical significance

Levene's test is based on the contrast of statistical significance [18].

By expert consensus, in the social and behavioral sciences, the significance level is commonly set at 5%, or 0.05 [19].

It is advisable to be conservative and adhere to this criterion.

Step 2. Formulate the test hypotheses

Based on the established significance level, the following test hypotheses can be set:

a) If $p > 0.05$, retain the null hypothesis (H_0).

b) If $p \leq 0.05$, reject the null hypothesis (H_0).

It is important to remember that the null hypothesis, which predicts no differences between the groups being measured, according to Kazdin [20], is typically formulated "with the express purpose of being rejected" [21].

However, in this case, the goal is to retain the null hypothesis to confirm equal variance between groups, which is essential for designating a control group in the study.

This consideration will be crucial when interpreting the results of the hypothesis test in Step 4.

Step 3. Conduct the Test Using SPSS

A practical procedure for calculating Levene's test is to follow the steps for performing an Independent Samples T-Test, as Levene's test is automatically generated as part of this analysis. The general steps are as follows:

1. In the "Variable View" tab, enter three variables:
    a) One variable corresponding to each participant in the study.
    b) Another variable representing the subject of study (the dependent variable).
    c) A third variable for the groups being compared (e.g., Experimental and Control). Ensure that the "Measure" for the group variable is set to Nominal.
2. In this same interface, label the Experimental and Control groups by assigning arbitrary numbers to each using the "Values" button (e.g., assign the number 1 to the Experimental group and 2 to the Control group).
3. In the "Data View" tab, enter the scores corresponding to the subject of study for each participant in both groups (these scores should have been collected beforehand). Also, enter the number assigned to each subject, indicating whether they belong to the Experimental or Control group, as established in Step 2.
4. Click the "Analyze" button. When the menu appears, select "Compare Means" and then choose "Independent-Samples T Test" from the submenu.
5. In the dialog box that appears, place the subject of study in the "Test Variable" field and the grouping variable (Experimental vs. Control) in the "Grouping Variable" field.
6. Click the "Define Groups" button and specify the values assigned to each group (Experimental and Control).
7. Finally, click "OK" to run the analysis and obtain the results.

Step 4. Interpret the Results

After obtaining the Independent Samples T-Test output, locate the section titled "Levene's Test for Equality of Variances". Compare the p-value (labeled as "Sig.") with the hypotheses established in Step 2.

a) If $p > 0.05$, this indicates that the groups have equal variances, meaning the Control group can be incorporated into the study.

b) If $p \leq 0.05$, this suggests that the groups are statistically different in their variances, indicating that the

Control group may not be suitable for inclusion, or adjustments must be made based on this inequality.

Graphical Representation of Levene's Test

Graphs are useful tools for reviewing and communicating the results of statistical tests by visually illustrating the behavior of the processed data. For Levene's test, Box-and-Whisker Plots can be generated in SPSS. These plots provide a clear visual representation of the distribution of scores within each group, allowing for an intuitive comparison of variance between the Experimental and Control groups.

General Process for Generating the Box-and-Whisker Plot in SPSS

1. Click on the "Graphs" option in the top toolbar of any SPSS editing window. From the dropdown menu, select "Legacy Dialogs", and in the following submenu, choose "Boxplot".
2. In the dialog box that appears, select the "Simple" option and click the "Define" button.
3. In the next window, place the subject or variable of study into the "Variable" field, and the grouping variable into the "Category Axis" field. Then, click "OK" to generate the plot.

Interpretation of the Box-and-Whisker Plot

The generated plot contains several elements that can be analyzed to understand the statistical behavior of the groups being compared. However, the element of interest for verifying homoscedasticity is the median.

The median is represented by the thick horizontal line inside each box or rectangle. To check for homoscedasticity, extend the line from the median in each box: if the extended line meets any point in the opposite box, the groups are considered homogeneous. If at least one of the medians does not meet this condition, homoscedasticity cannot be confirmed.

An example case is illustrated in Figure 2:

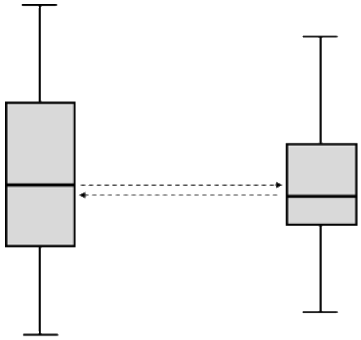

Figure 2. Verification of coincidence of the median between the compared boxes

As observed in the example, the medians of both boxes, when extended towards each other, almost coincide exactly, indicating homogeneity between the groups. It is important to clarify that this proximity—almost exact—of the medians in each group is not a mandatory condition for denoting homoscedasticity. As mentioned, it is sufficient for the coincidence to occur at any point of the opposing box in both cases.

In addition to this, the box-and-whisker plot can illustrate other outcomes or behaviors of the data within the groups, such as the concentration or dispersion of scores, the interquartile range, symmetry, lower and upper limits, outliers, and extreme values, among others. These elements can provide valuable insights; however, they will not be addressed in this work for reasons of space and purpose.

3. Conclusions

Having information about the homogeneity of variances among the groups being compared enables researchers to make methodical decisions with a more solid technical foundation. These decisions may lead to one of the following scenarios:

Confirming the equality of variances allows for the comparison of groups in a quasi-experiment, which, in turn, enhances the robustness of the research method [22].

However, this procedure does not guarantee greater control over the variable or variables involved in the study

groups; rather, it serves to justify the decision to subject these groups to comparison. For example, one of the most notable limitations of the test is its failure to account for linear dependencies among the resulting residuals [23].

Therefore, it is necessary to emphasize the importance of thorough planning for the activities to be conducted in the study. Once the technical determination of a control group has been established, historical cohorts can serve as a viable option to enhance the robustness of this procedure and mitigate some of the threats to internal validity in quasi-experimental studies [24].

Conversely, when there is no evidence of homogeneity of variances between the experimental group and the proposed control group, the most prudent course may be, at first, to opt for a pre-experimental design that omits a control group. However, this approach would consequently increase the necessity for meticulous planning in the research process, as this is the model with the least control among the types of experiments [25].

Accordingly, this design can adequately address research objects that theory suggests are not easily influenced by maturation effects.

Alternatively, retaining the control group even when the Levene test indicates statistical significance, that is, heterogeneity of variances, will necessitate the use of non-parametric tests to address the research hypothesis.

References:

[1] Kerlinger, F. N. and Lee, H. B. (2002). “Investigación del comportamiento. Métodos de investigación en ciencias sociales”.McGraw Hill.

[2] Mai, N. N., Takahashi, Y. and Oo, M. M. (2020). “Testing the Effectiveness of Transfer Interventions Using Solomon Four-Group Designs”. https://www.mdpi.com/2227-7102/10/4/92

[3] Solomon, R. L. (1949). “An extension of control group design”. Psychological Bulletin, 46(2), pp.137–150. https://psycnet.apa.org/record/1949-05862-001

[4] Namakforoosh, M. N. (2014). “Metodología de la investigación”. Limusa.

[5] White, H. and Sabarwal, S. (2014). “Quasi-Experimental Design and Methods”. UNICEF. https://www.unicef-irc.org/publications/pdf/brief_8_quasi-experimental%20design_eng.pdf

[6] Thorndike, E. L. (1904). “An introduction to the theory of mental and social measurements”. Ulan Press.

[7] Tornimbeni, S., Pérez, E. and Olaz, F. (2014). “Introducción a la psicometría”. Paidós.

[8] Landero, R. and González, M. T. (2006). “Estadística con SPSS y metodología de la investigación”. Trillas: Universidad Autónoma de Nuevo León.

[9] Creswell, J. W. (2013). “Research design. Qualitative, Quantitative and Mixed methods approach”. Sage Publications Inc.

[10] Moreno, M. G. (2000). “Introducción a la metodología de la investigación educativa II”. Progreso.

[11] Bunge, M. (1983). “La investigación científica”. Ariel Methodos.

[12] Muñoz, C. I. (2015). “Metodología de la investigación”. Oxford.

[13] Arias, F. (2014). “Metodología de la investigación”. Trillas.

[14] Starkweather, J. (2022). “Homogeneity of Variances”. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/Benchmarks/Levene_JDS_Mar2010.pdf

[15] Nordstokke, D. W. and Zumbo, B. D. (2010). “A New Nonparametric Levene Test for Equal Variances”. Psicológica, 31(2), 401-430. https://www.redalyc.org/pdf/169/16917017011.pdf

[16] Nordstokke, D. W., Zumbo, B. D., Cairns, S. L. and Saklofske, D. H. (2011). “The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data”. Practical assessment, Research and Evaluation. Vol 16 (Num 5).

[17] De Almeida, A., Elian, S. and Nobre J. (2008). “Modifications and Alternatives to the Tests of Levene and Brown & Forsythe for Equality of Variances and Means”. Revista Colombiana de Estadística. (2)31, pp. 241-260.

[18] Rodríguez, C., Gutiérrez, C. and Pozo, T. (s. f.). “Fundamentos conceptuales de las principales pruebas de la significación estadística en educación”. Grupo Editorial Universitario.

[19] Hinkle, D. E., Wiersma, W. and Jurs, S. G. (2003). “Applied Statistics for the Behavioral Sciences” (5th ed.). New York: Houghton Mifflin Company.

[20] Marczyk, G., DeMatteo, D. and Festinger, D. (2005). "Essentials of research design and methodology". John Wiley & Sons, Inc.

[21] Siegel, S. and Castellan, N. (1995). "Estadística no paramétrica aplicada a las ciencias de la conducta". Trillas.

[22] Nordstokke, D. W. and Colp, S. M. (2014). "Investigating the robustness of the nonparametric Levene test with more than two groups". Psicológica, 35. https://www.uv.es/revispsi/articulos2.14/10NORDSTOKKE.pdf

[23] Brown, M. B. and Forsythe, A. B. (1974). "Robust tests for the equality of variances". Journal of the American Statistical Association.

[24] Walser, T. M. (2014). "Quasi-Experiments in Schools: The Case for Historical Cohort Control Groups". North Carolina University. ERIC number: EJ1031270

[25] Nordstokke, D. W. and Zumbo, B. D. (2007). "A Cautionary Tale About Levene's Tests for Equal Variances". https://files.eric.ed.gov/fulltext/EJ809430.pdf