brazilianpoliticalsciencereview

FORUM

Comments on When is Statistical Significance not Significant?*

Glauco Peres da Silva
Universidade de São Paulo, Brazil

Fernando Henrique Guarnieri
Centro de Estudos da Metrópole, Brazil

(FIGUEIREDO FILHO, Dalson. B. et al. When is statistical significance not significant? *Brazilian Political Science Review*. Vol. 07, N° 01, 2013)

An article titled "*When is statistical significance not significant?*" was published in this journal. The study echoes countless others which have underlined the dangers of wrongly interpreting the p-value (see Gill (2004) for an overview of such studies). The authors' main goal is to provide a basis for understanding the logic behind tests of significance and their appropriate role in empirical studies within the social sciences. For this, the authors present four recommendations with universalistic intents: 1) to always graphically analyze data before interpreting the p-value; 2) it is pointless to estimate the p-value for non-random samples; 3) the p-value is highly affected by the sample size; and 4) it is pointless to estimate the p-value when dealing with data on population.

We believe that these recommendations do not achieve the intended objectives, even when dealing with very simple examples. This is problematic since the authors intend "to help students to make sense of the appropriate role of the *p-value* statistics in empirical political science research" (p. 32). Given that these suggestions orient the definition of the mathematical relationship among variables, which in turn is a result of theoretical constructs, we intend to show how distant we still are when we follow the suggested steps. Therefore, our motivation is to indicate points unmet by the proposed recommendations. The aim is also to collaborate with the practical use of statistical analysis in political science and thus complementing the discussions raised by the authors. It is worth mentioning that we generally agree with the assessments on the limits of using the *p-valu*e, but we intend to open here an opportunity to discuss the recommendations followed from these diagnoses.

Recommendation: Scholars must always graphically analyze their data before interpreting the *p-value.*

Graphical analysis is useful and powerful diagnostic tool for linear relationships without the presence of covariates. In this case a scatter graph would show a pattern of relationship

(*) http://dx.doi.org/10.1590/1981-38212014000100018
The R simulation code used in this article can be found in bpsr.org.br/files/arquivos/Banco_Dados_Peres_Guarnieri.rtf.
This Research was supported by FAPESP. Process: 2009/14768-2.

between the two variables to help the researcher. This is an exceptional case, however. One could consider any number of situations where graphical analysis would not help the researcher since it depends on the relationship between the considered variables, which in turn depends on theoretical keystones, the importance of covariates in the models, and the functional form used. The authors only briefly mention this aspect by saying that "[the] practical consequence of functional form misspecification in this case is the underestimation of the magnitude of relationship between variables." (p. 39).

A familiar example extracted from Taagepera (2012) is quite illustrative of this point: the author calls for other researchers to estimate the non-linear relationship represented by the universal gravitation formula. The result, according to the author, is that an erroneous relation is identified both due to lack of a theory that allows one to find what they seek within the data and because one stemmed from a generally linear relationship between variables. A simulation with this data shows that the graphical evaluation would not be enough to avoid a misguided analysis. That is, reinforcing Taagepera's argument on which we agree, in order to find the adequate functional relationship, researchers must theoretically know what they aim to find in the data before evaluating the p-value or even carrying out a graphical analysis. This would be secondary to a functional form analysis derived from a theory supporting it.

Recommendation: It is pointless to estimate the *p-value* for non-random samples

This is simply not true. The *p-value*, as conceived by Fisher (1925) is a measure for the adjustment of a model or "null hypothesis" to our data. Conditional upon the null hypothesis being true, if the observed values discern enough from the expected values we can then reject the hypothesis that the difference between them is due to chance. For example, if a study aims to compare different data sets, the sample size may not be random and the *p-value* could still be useful to distinguish an interesting characteristic. This is a common situation in political science. A fairly trivial example is to assess whether the deputies who comprise the governing coalition during a certain president's term are more disciplined than in previous legislatures. The samples considered in this situation are not randomly produced and the *p-value* is helpful to provide the difference between these groups and the knowledge produced through this research is of interest to the field.

Recommendation: The *p-value* is highly affected by the sample size

The authors suggest that we can always obtain a significant *p-value* if we increase our sample size thus indicating a distinction between the power of the test and the significance of the test, something only marginally evaluated by the authors. To support this argument they simulate the power of a significance test with different sample sizes. For large samples the test power is 100%. Does this mean that if you increase your sample size you will get a significant *p-value*? Not necessarily. The power of the test provides us with the probability of rejecting a null hypothesis when it is false. Should the null hypothesis be false we will certainly reject it from a given sample size. But this does not mean that we will reject it if the null hypothesis is true. Remember, the *p-value* provides a starting point by which we reject the null hypothesis on the condition that is true. We may continue to obtain non-significant p-values even when our test power is 100%. Evidently, when considering the very mathematical formulation of Student's t test, the larger the sample size, the higher the *p-value*. But this does not lessen the importance of differentiating the power of the test from its significance, which seems to be a more central issue for empirical studies test.

As the authors rightly point out, effects of a small magnitude (from a substantive point

of view) may become significant with a larger sample, though they would still continue to be of a small magnitude. This is not because of the power of the test, but due to the fact that a larger sample almost always provides us with more information, similar to observing space with a more powerful telescope: more information is always good. Even in situations where we are unable to increase our sample, we may obtain significant results and powerful tests with small samples. Lastly, effects with significant p-values on large samples are just as significant as significant *p-values* in small samples.

Recommendation: It is pointless to estimate the p-value when dealing with data on population

As scientists we are interested in describing and explaining phenomena. The explanation strives to be as general as it can. This means that every time a theoretically determined mechanism is present under certain circumstances a certain predicted outcome by the theory shall occur. We check our theories against our data. Positively, each new observation gives us a new test to our theory and each time we fail to reject it we have more confidence in its validity. The data obtained by any applied technique are derivations of theoretical constructions upon the empirical world, our concepts. They therefore help us to understand the abstract connections built upon their concrete dimensions. In this sense even when we came across a census it is but one of many instances in which we test our hypothesis. In this sense it stands to reason to think in terms of *p-value*. Even in the task of merely describing something the *p-value* makes sense, since even a census may contain flaws, such as measurement error.

General Considerations

Given these four recommendations the authors conclude that when a researcher uses significance tests, they must ensure they are working with a large random N sample. What we said above changes this final recommendation. The *p-value* may be useful even if our sample is not random nor as large as statistics books recommend (for a good example of this as well as examples where the *p-value* could hinder the researcher see Gelman, 2011).

In the conclusion the authors call upon researchers not to confuse a significant *p-value* effect with a significant theoretical and substantive effect. We most certainly agree with such a recommendation[1]. We also agree that the lack of statistical significance does not mean a less important finding from a theoretical point of view. The difference between a significant result and a non-significant result is not necessarily in itself significant. The difference between a study with an average of 25 and standard error of 10 (therefore significant to a *p-value* of 0.01) and another study with an average of 10 and 10 standard errors (not significant) have an average of 15 and an error of 14 and is therefore not significant[2].

But the solution to these issues is not to analyze the data graphically and to work with large random N samples. Our recommendation is that researchers who wish to avoid the problems of the p-value should always remember that it always gives us the probability that

1 Cinelli (2012) carries out a similar analysis in the Economy field.

2 It is interesting to note that this does not depend either on substantive issues or on the sample size, since in order to calculate the confidence interval and the standard deviation of the difference of means only two pieces of information are deemed necessary: means and standard errors. Remembering that the calculation of the confidence interval at 95% occurs by the following equation: mean±1,96 (SE), where SE is the standard error, and also, in order to find to find the standard error in a difference of means we have σ12n1+σ22n2 , where σj2 is the variance in each sample. Since SE= σn, where σ is the standard deviation, we have that SE= σ2n. Therefore, the difference of means can be rewritten as follows: SE12+SE22 (example taken from Gelman and Stern (2006)).

our results are due to the conditional chance that the null hypothesis is true (P(D|Ho)). It is not a test to verify if Ho is true (P(Ho|D)). Once we do not reject the null hypothesis this does not mean that our data was caused by chance or that there is no connection between the variables. Empirical work is closely tied with the theoretical questions behind it and it should be conducted in their light.

Following Gill's footsteps (1999) we also recommend that instead of merely reporting the p-values, researchers should attempt to flee from the perils of the hypothesis test and further explore their findings to the limit. Limiting an analysis to the level of significance of certain effects is a poor theoretical practice. We can report effects observed by means of techniques such as predicted values, expected values, first differences (KING et al., 1998) for assessing a theory in terms of the observed results and not only by the presence of a certain effect; we may use simulations to obtain the distribution of a conditional parameter relevant to our data and thereby test counterfactuals (GELMAN and RUBIN, 1995); or we can simply report the confidence interval which expresses our uncertainty as to our estimation and provides us with the same kind of information as the p-value.

We hereby believe to have presented consistent arguments for the case that these four given recommendations are also not free of controversy. Perhaps the most general recommendation is that researchers who plan on using statistical analyses techniques should pay attention to the implications behind this methodological relationship in face of the desired theoretical results so that relevant questions may be answered. That is to say, when a scientist wishes to know something, they use a certain set of established procedures. But these should not be merely taken as dogmas. There are limits to them and knowing such limits is vital.

Translated by Paulo Scarpa
Submitted in August 2013
Accepted in April 2014

References

CINELLI, Carlos L. K. (2012), Inferência estatística e a prática econômica no Brasil: os (ab)usos dos testes de significância. Dissertação de Mestrado, Departamento de Economia, Universidade de Brasília.

GELMAN, Andrew and RUBIN, Donald (1995), Avoiding model selection in Bayesian social research. Sociological Methodology, nº 25, pp.165-174.

GELMAN, Andrew and STERN, Hal (2006), The difference between 'significant' and 'not significant' is not itself statistically significant. The American Statistician, Vol. 60, nº 04, pp. 328-331.

GELMAN, Andrew (2011), Why tables are really much better than graphs. Journal of Computational and Graphical Statistics. Vol. 20, nº 01, pp. 03-07.

GILL, Jeff (1999), The Insignificance of Null Hypothesis Significance Testing. Political Research Quarterly, vol. 52, nº 3, pp. 647-674.

GILL, Jeff (2004), "The Current Paradigm: Null Hypothesis Significance Testing", in http://www.stats.org.uk/statistical-inference/Gill.pdf

KING, Gary; TOMZ, Michael; WITTENBERG, Jason (1998), How to Interpret and Present Statistical Results or Enough with the Logit Coefficients Already! In: Annual meetings of the American Political Science Association.

TAAGEPERA, Rein (2012), Logical Models and Basic Numeracy in Social Sciences, in http://www.psych.ut.ee/stk/Begginers_Logical_Models.pdf