European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available online at: www.oapub.org/edu

DOI: 10.46827/ejes.v11i12.5712 Volume 11 | Issue 12 | 2024

SECOND-GENERATION P-VALUES: A NEW PRACTICAL APPROACH FOR EDUCATIONAL RESEARCH

Sally A. Lesik[i],
Anna M. Duffin
Central Connecticut State University,
USA

Abstract:
The limitations, misconceptions, and misuse of traditional p-values remain a concern for researchers in education. Second-generation p-values provide a descriptive measure that can be used to decide between the null and alternative hypotheses. They are intuitive and easy to understand and provide a new approach to addressing some of the limitations of traditional p-values, mostly by considering the practical implication of an estimated effect. This article describes the general idea and principles behind second-generation p-values and illustrates their application in educational research and data science. Limitations and directions for further study are discussed.

Keywords: statistical inference; p-value; second-generation p-value; interval null hypothesis; effect size; practical significance; statistical significance

1. Introduction

Over the past decade, there has been growing controversy regarding how the results of statistical methods are reported across the body of scientific literature (Campbell et al., 2019). Perhaps the biggest controversy is how the p-value is used and misused (Kmetz, 2019). In 2016, the American Statistical Association (ASA) released a pivotal statement to address some of the misconceptions and misuse of the traditional p-value (Wasserstein and Lazar, 2016). This statement addresses how the p-value by itself does not provide a good measure of evidence, nor does it measure the effect size, and that scientific conclusions and education policy decisions should not be based solely on whether the p-value is less than a given threshold. Many applied researchers have been investigating how the p-value is used within their discipline and provide ideas of how to improve clarity and transparency (Di Leo and Sardanelli, 2020; Goodman, 2008; Greenland et al., 2016).

[i] Correspondence: email lesiks@ccsu.edu

It is customary to focus exclusively on research with p-values that describe "statistical significance" (Hurlbert et al., 2019). Discussing methods or practices lacking such significance is rare. Often, data reporting and research questions are skewed towards statistically significant outcomes merely to meet publication criteria. Some researchers even admit to desperately seeking statistical significance, driven by the belief that research lacks meaning without it (Robert, 2020). It is important to note that obtaining non-statistically significant test results does not necessarily render the research meaningless or devoid of valuable insights. It could serve as a starting point for further exploration, potentially leading to new avenues of inquiry, stimulating meaningful discussions and opening avenues to unexpected patterns. Shifting the approach is essential to have opportunities for insightful conclusions. The conventional misuse of p-values while trying to conform to traditional norms ultimately compromises the integrity of research and limits opportunities for meaningful discourse and comprehensive analyses and outcomes (Colquhoun, 2017; Gao, 2020). This is just one reason why many argue to abandon the term "statistical significance" altogether (Diaz-Quijano et al., 2020; McShane et al., 2019; Wasserstein et al., 2019).

Perhaps the most notable misconception about traditional p-values is when a p-value is less than some threshold, that this automatically implies significance or importance. This misconception can lead to the belief that if a research result is statistically significant, then it leads to that result being practically meaningful. The practicality of the results needs to be considered, as well as the effect size and the significance of the findings. For instance, if a new teaching method is being evaluated, it is crucial to look at the effect size, practical relevance of the findings, and the context of the new method or intervention. Making conclusions solely based on a p-value by itself can lead to misconceptions and incorrect interpretations.

Even as far back as the nineteenth century, Edgeworth's (1885) description of using the p-value as a tool for signaling results is worth closer examination. It is important to note that this concept was never intended to equate statistical significance with practical or clinical importance. It was further observed that there is no statistical importance in distinguishing between 'significant' and 'not significant' findings. However, this misconception could falsely imply results are worth attention, which could lead to selective reporting (Vidgen B and Yasseri T, 2016).

P-values are typically reported based on a point null hypothesis. One of the major drawbacks of a point null hypothesis is that because it is an exact value, it can never be accepted. The basic idea with hypothesis testing is to reject a point null hypothesis when the sample data appears to be drawn from a population whose parameter of interest appears to be different enough from the given point estimate expressed in the null hypothesis. In other words, the null hypothesis is rejected, and the alternative hypothesis is accepted. Statistical significance is often described using a p-value, which corresponds to the observed level of significance that is based on a test statistic, and it is the lowest level of significance for which the null hypothesis can be rejected. If a p-value is less than a given threshold (which is typically 0.05 in most educational research), this leads to accepting the alternative hypothesis and rejecting the null hypothesis.

However, it is possible to achieve statistical significance even if an effect (or difference) does not have any practical meaning or clinical impact. There can also be an effect (or difference) that does not reach statistical significance, but there is a meaningful practical impact. A practical impact is when the effect has meaning in the real world, in other words, when the effects (or differences) are large enough to be of practical or clinical value.

To illustrate this concept, suppose the effectiveness of a new educational intervention program is being evaluated. A random assignment was done where one group was exposed to the intervention program (I), and another group was not (N). Both groups were then given an assessment on a scale from 0 to 100. Table 1 gives the sample size, mean, and standard deviation of the assessment score for a hypothetical scenario for 6,000 students from each of the two different groups.

Table 1: Descriptive statistics for the assessment scores for the sample of 6,000 students in the intervention group and 6,000 students in the non-intervention group

Intervention (I)	**Non-Intervention (N)**
n = 6,000	n = 6,000
$\bar{x} = 75$	$\bar{x} = 74.7$
s = 5	s = 5

Running a two-sample t-test ($H_0: \mu_I = \mu_N, H_A: \mu_I \neq \mu_N$), the estimated mean difference between the two groups generates a p-value less than 0.05 ($p \approx 0.001$), which is "statistically significant", showing an improvement among students who received the intervention compared to those who did not participate in the intervention. However, the estimated difference of 0.3 points on the assessment measure does not have much of a practical impact at all on a scale from 0 to 100. Clearly, larger samples have smaller standard deviations and will find smaller effect sizes, even if there is no practical or clinical impact.

Further analyzing this example, the significant p-value suggests that the observed improvement is unlikely to have occurred by chance alone. However, the analysis indicates a small estimated effect of only 0.3 points, suggesting that the effect of the intervention might not be practically meaningful. This miniscule effect size translates into only a modest increase in the achievement score, which does not seem to appear to have a substantial impact on students' overall performance. The intervention implementation might have led to a statistically significant improvement in achievement, but the required resources, financial or logistical, may be too large for such a modest increase. In the context of researching this new intervention, where the achievement score shows no meaningful improvement over the current method, conversations around the efficacy of the current method and the rationale for considering a change hold more value than simply accepting inflated claims of the superiority of the intervention.

Now consider another scenario where Table 2 provides the sample size, mean, and standard deviation of the assessment score for a hypothetical example of 5 students from the two different groups.

Table 2: Descriptive statistics of the assessment scores for the sample of 5 students in the intervention group and 5 students in the non-intervention group

Intervention (I)	Non-Intervention (N)
$n = 5$	$n = 5$
$\bar{x} = 75$	$\bar{x} = 65$
$s = 15$	$s = 15$

While the 10-point difference between the two groups does not reach statistical significance as the p-value is greater than 0.05 ($p \approx 0.323$), it clearly does have a practical impact, equating to an improvement of an entire letter grade (from a D to a C). Smaller sample sizes tend to have more variability (larger standard deviations) so it can be difficult to find even a large practical effect.

Finally, consider another scenario where Table 3 provides the sample size, mean, and standard deviation of the assessment score for a hypothetical example of 85 students from the two different groups.

Table 3: Descriptive statistics of the assessment scores for 85 students in the intervention group and 85 students in the non-intervention group

Intervention (I)	Non-Intervention (N)
$n = 85$	$n = 85$
$\bar{x} = 75$	$\bar{x} = 68$
$s = 10$	$s = 10$

For this scenario, the 7-point difference reaches statistical significance ($p \approx 0.000$) and also has a practical impact, equating to an average improvement for those exposed to the intervention of more than half of a letter grade.

2. Second-Generation p-Values

There has always been, and will always be, a need for rapid data assessment in educational research. However, it is crucial to understand that decisions cannot solely rely on traditional p-values. This is where second-generation p-values (SGPVs) can come into play.

SGPVs are a descriptive statistic that can be used to describe the proportion (or percent) of the data that are consistent with an interval null hypothesis versus a probability statement as with the traditional p-value. It can be described as the proportion of a confidence interval that overlaps a specified interval null hypothesis. An interval null hypothesis consists of the range for any trivial effects. SGPVs are not designed to replace traditional p-values but rather to complement and enhance them. They can indicate when the null hypothesis is supported or when there is significant evidence for the alternative hypothesis (Blume et al., 2019).

Researchers do not need extensive statistical training to utilize SGPVs, as they offer more intuitive and interpretable results when compared to traditional methods. This allows for a flexible interpretation of important findings, granting researchers the

freedom to holistically consider the available data and make judgments based on the strength of evidence provided rather than strictly relying on a rigid threshold for statistical significance. One advantage of SGPVs is that they are easy to calculate and interpret. Furthermore, they can also be incorporated within frequentist, Bayesian, or Likelihood analyses, and Type I and Type II error rates converge to 0 (Stewart et al., 2019). In today's data-rich world, SGPVs also have excellent large-sample properties (Stahel, 2021). And unlike traditional p-values, SGPVs eliminate the possibility of p-value hacking and selective reporting to fit the criteria of statistical significance (Das and Das, 2023; Vidgen and Yasseri, 2016), and, therefore, provide a clearer and more reliable approach

Essentially, SGPVs can be used with confidence intervals for any statistical method. And most notably, SGPVs incorporate the practical or clinical significance of an estimated effect.

3. Methodological Details

Let θ be an unknown population parameter and let $I = \left(\hat{\theta}_L, \hat{\theta}_U\right)$ be a 100(1- α) % confidence interval estimate of θ with length $|I| = \left|\hat{\theta}_U - \hat{\theta}_L\right|$, where $\hat{\theta}_L$ is the lower limit and $\hat{\theta}_U$ of is the upper limit of the confidence interval . If H_0 represents the null hypothesis expressed in the form of an interval, $H_{0:}\, \theta \pm \delta$, with length $|H_0| = |2\delta|$, then the SGPV, (denoted as p_δ) can be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

Where $|I \cap H_0|$ is the amount of overlap between the confidence interval and the interval null hypothesis.

There are three conclusions that can be obtained with SGPVs. A SGPV can either support the alternative hypothesis by indicating that the entire confidence interval lies outside of the null interval. This is when $p_\delta = 0$, which indicates that 0% of the data is compatible with the null hypothesis. It can also support the null hypothesis when the entire confidence interval lies within the null interval. This is when $p_\delta = 1$, which indicates that 100% of the data is compatible with the null hypothesis. SGPVs can also be inconclusive and this occurs when $0 < p_\delta < 1$.

The small sample correction factor, $\left(\frac{|I|}{2|H_0|}\right)$, is typically used for small samples when the power drops below 16% (Blume et al., 2018).

3.1 Examples

Returning to the hypothetical study to compare the difference in achievement for two groups, one group exposed to an intervention and another that was not. By considering the first scenario (Table 1), and assuming that anything within a 3-point difference is not practically meaningful, then the null interval can be expressed as H_0 = (-3, 3) where, δ =

3, with length $|H_0| = |2\delta| = |2 \cdot 3| = 6$. A 95% confidence interval for the difference between the two population means would be I = (0.121, 0.479) with length $|I| \approx |0.479 - 0.121| = 0.358$. Then the SGPV would be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \frac{0.358}{0.358} \cdot max\left\{\frac{0.358}{2(6)}, 1\right\} = \frac{0.358}{0.358} \cdot 1 = 1$$

Notice that the confidence interval I = (0.121, 0.479) is entirely contained within the null interval H_0 = (-3, 3), and so $I \cap H_0 = I$ with length $|I \cap H_0| = |I|$. And since $p_\delta = 1$, this supports the null hypothesis, as 100% of the confidence interval overlaps the interval null hypothesis, so there is no meaningful effect.

Now considering the second scenario (Table 2), and similar to the previous example, assuming that anything less than a 3-point difference is not practically meaningful, $|H_0| = 6$. A 95% confidence interval for the difference in the two-population means is I = (-11.88, 31.88) with length $|I| \approx |31.88 - (-11.88)| = 43.76$. The SGPV would be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \frac{6}{43.76} \cdot max\left\{\frac{43.76}{2(6)}, 1\right\} = \frac{6}{43.76} \cdot \frac{43.76}{2(6)} = \frac{1}{2}$$

Because the null interval H_0 = (-3, 3) is contained entirely within the confidence interval I = (-11.88, 31.88), then $|I \cap H_0| = |H_0|$. For this scenario, the small sample correction factor $\frac{|I|}{2|H_0|}$ is used, given the small sample size. And since $p_\delta = 1/2$, the results are inconclusive.

For the third scenario (Table 3), and again, assuming anything less than a 3-point difference is not practically meaningful, $|H_0| = 6$. A 95% confidence interval for the difference in population means would be I = (3.97, 10.03) with length $|I| \approx |10.03 - 3.97| = 6.06$. The SGPV would be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \frac{0}{6.06} \cdot max\left\{\frac{6.06}{2(6)}, 1\right\} = \frac{0}{6.06} \cdot 1 = 0$$

Notice that the null interval H_0 = (-3, 3) and the confidence interval I = (3.97, 10.03) do not overlap, so $|I \cap H_0| = 0$. And since $p_\delta = 0$, this supports the alternative hypothesis as there is 0% overlap between the interval null hypothesis and the corresponding confidence interval. The effect is statistically significant ($p \approx 0.000$) and it is also practically meaningful based on the predetermined range of the interval null hypothesis.

3.2 A Natural Connection to Large Data Sets

One area where SGPVs can have a notable impact is in the field of data science. Since data science relies on large data sets with massive amounts of information, p-values quickly

converge to 0, even for minuscule effect sizes that have little or no practical meaning (Lin et al., 2013).

There are many large data sets related to education that are available. One example are the data sets from the 2019 and 2021 National Survey of College Graduates (https://ncsesdata.nsf.gov/sestat/). These data sets provide detailed information about the employment, educational experience, and demographic characteristics for a large sample of recent college graduates at two different points in time: 2019 which was pre-COVID, and 2021, which was post-COVID.

Two questions were asked in both of these surveys: (CCCOLPR), whether a student responded that they attended a community college to prepare for college or to increase the chance of being accepted to a 4-year college or university, and (CCFIN), whether a student responded that they attended a community college due to financial reasons. The number of students who responded as "Yes" or "No" to these two different questions is provided in Table 4 and Table 5, respectively.

Table 4: CCCOLPR, Number of Yes and No responses for 2019 and 2021

Response	**Year**	
	2019	**2021**
Yes	13,559	16,058
No	24,374	26,592
Total	**37,933**	**42,650**

Table 5: CCFIN, Number of Yes and No responses for 2019 and 2021

Response	**Year**	
	2019	**2021**
Yes	14,743	17,150
No	23,190	25,500
Total	**37,933**	**42,650**

Suppose a researcher is interested in estimating the difference between the proportion (or percentage) of students who responded Yes to attending a community college to increase the chance of getting into a 4-year college or university (CCCOLPR) between the years 2019 (pre-COVID) and 2021 (post-COVID). If the null interval represents any percent difference of less than 2% as being trivial, then the null interval for the proportion would be represented as H_0 = (-0.02, 0.02) where $|H_0| = 0.04$. The respective 95% confidence interval for the difference in the two proportions using the normal approximation would be I = (-0.0257, -0.0124), with length $I = |-0.0124 - (-0.0257)| = 0.0133$. Then the SGPV would be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \frac{0.0076}{0.0133} \cdot max\left\{\frac{0.0133}{2(0.04)}, 1\right\} \approx 0.57 \cdot 1 = 0.57$$

Notice that $I \cap H_0 = (-0.02, -0.0124)$, with length $|I| \approx |-0.0124 - (-0.02)| = 0.0076$. Since $p_\delta = 0.57$, this is inconclusive as there is a 57% overlap between the

confidence interval and the null interval. Given the large sample size, it is no surprise that the p-value for the difference in proportions ($H_A: p_{2019} \neq p_{2021}$) is significantly less than any reasonable level of significance, in fact, $p \approx 2.07 \times 10^{-8}$, even though the difference does not have a practical impact based on the specification of the interval null hypothesis. Therefore, based on the SGPV, there is no meaningful difference in the proportion of students who attended a community college to increase the chance of getting into a 4-year college or university between 2019 (pre-COVID) and 2021 (post-COVID).

Similarly, if a researcher is interested in estimating if there is a difference between 2019 and 2021 in the proportion (or percent) of students who responded Yes to attending a community college for financial reasons (CCFIN), and suppose that the null interval represents any percent difference less than 0.5% as being trivial.

The null interval for the proportion would be H_0 = (-0.005, 0.005) where $|H_0| = 0.010$. The respective 95% confidence interval for the difference in the two proportions using the normal approximation would be I = (-0.0202, -0.0067) with length $|I| = 0.0135$. Then, the SGPV would be calculated as follows:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \cdot max\left\{\frac{|I|}{2|H_0|}, 1\right\} = \frac{0}{0.0135} \cdot max\left\{\frac{0.0135}{2(0.01)}, 1\right\} = 0 \cdot 1 = 0$$

Notice that the confidence interval and the null interval do not overlap, so $|I \cap H_0| = 0$. Since $p_\delta = 0$, the results are compatible with the alternative hypothesis as there is a 0% overlap between the confidence interval and the null interval. Given the large sample size, it is no surprise that the p-value for the difference between the two proportions ($H_A: p_{2019} \neq p_{2021}$) is significantly less than any reasonable level of significance ($p \approx 9.65 \times 10^{-5}$). Given the range of trivial differences that is specified in the interval null hypothesis, the effect is practically meaningful. Therefore, based on the SGPV, there is approximately a 0.5% difference in the percent of students who attended a community college for financial reasons between 2019 (pre-COVID) and 2021 (post-COVID).

Clearly, it is the specification of the range of the interval null hypothesis that is crucial in determining if there are meaningful effect sizes as the p-values will tend to converge to 0 for large sample sizes, thus almost always suggesting statistical significance even if there is no practical or clinical impact.

4. Summary

SGPVs add transparency to research findings as well as their interpretation by providing additional information about the effect size, uncertainty, and strength of evidence. It fosters greater confidence in research findings. And while SGPVs are very easy to calculate manually, there are statistical software packages, such as Stata (Bormann, 2022) and R (Package "Sgpv", 2022; Zuo et al., 2022) that calculate SGPVs and allow for visualization with features such as interactive graphs and customizable outputs.

Perhaps the greatest challenge in using SGPVs is specifying an interval null hypothesis versus a point null hypothesis prior to the onset of data collection. This means that the researcher needs to have a valid estimate of the magnitude and direction of trivial effects before data collection begins. And while this appears to be a new perspective, the practice of reporting confidence intervals and abandoning p-values or lowering the threshold of p-values is not at all uncommon (i.e. Armstrong and Hubbard, 2008; Cohen, 2021; Pericchi et al., 2014). Another consideration is that the state of applied research is to focus on finding a difference or an effect, in other words accepting an alternative hypothesis. This leads to a natural aversion to null or trivial effects, whereas by using SGPVs, this will incorporate a new dimension to finding meaningful effects as the null or trivial effects are explicitly described.

Just as with traditional p-values, it is always possible to manipulate data and analyses, and SGPVs are not much different. SGPVs can be manipulated by changing the width of the null interval until a desired result is achieved. However, by having established a predetermined null interval before data collection begins, this could help to eliminate any such issues.

Furthermore, when a SGPV is inconclusive $(0 < p_\delta < 1)$, currently there is no formal method that can be used to decide if a SGPV is close to 0 that would imply consistency with the alternative hypothesis, or if a SGPV is close to 1, that this could suggest consistency with the null hypothesis. For instance, if $p_\delta = 0.049$, could this imply consistency with the alternative hypothesis, as there is only a 4.9% overlap between the interval null hypothesis and the confidence interval? Or if $p_\delta = 0.94$, could this suggest consistency with the null hypothesis as there is a 94% overlap between the interval null hypothesis and the confidence interval? More research is needed in this area.

One strategy that can be used to address using the traditional p-value by itself, as is mentioned in the ASA's statement on p-values (Wasserstein and Lazar, 2016), would be to report SGPVs in addition to traditional p-values. By reporting both traditional p-values and SGPVs, scientific conclusions and policy decisions would not be based solely on whether the p-value is less than some predetermined threshold, as it would also support establishing meaningful effect sizes. And while large data sets have more of a presence in applied educational research, including SGPVs are clearly a better alternative as compared to reporting traditional p-values alone.

Conflict of Interest Statement

The authors declare no conflicts of interest.

About the Author(s)

Sally A. Lesik is a Professor of Mathematical Sciences at Central Connecticut State University where she teaches undergraduate and graduate courses in mathematics, statistics, and data science. Dr. Lesik's research agenda is focused on applied statistics, mathematics and statistics education, and policy evaluation in higher education.

Anna M. Duffin is an Assistant Professor of Mathematical Sciences at Central Connecticut State University, where she teaches undergraduate courses in mathematics, mathematics education, and statistics. Her research interests focus on mathematics education, statistics education, curriculum development, and strategies for supporting multilingual learners in the mathematics classroom.

References

Armstrong, J.S. and Hubbard, R. (2008). Why we don't really know what "statistical significance" means: A major educational failure, *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.1154386.

Blume, J.D., Greevy, R.A., Welty, V.F., Smith, J.R. and Dupont, W.D. (2019). An Introduction to Second-Generation p-Values. *The American Statistician*, 73(sup1), pp.157–167. doi: https://doi.org/10.1080/00031305.2018.1537893.

Blume, J.D, D'Agostino McGowan, L, Dupont, W.D, Greevy, R.A., (2018) Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. PLoS ONE 13(3): e0188299. https://doi.org/10.1371/journal.pone.0188299.

Bormann, S.-K. (2022). 'A Stata implementation of second-generation P-values', *The Stata Journal: Promoting communications on statistics and Stata*, 22(3), pp. 496–520. doi:10.1177/1536867x221124466.

Campbell, H. and Gustafson, P. (2019). 'The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication', *The American Statistician*, 73(sup1), pp. 358–373. doi: 10.1080/00031305.2018.1555101.

Cohen, M.P. (2021). 'Why not an interval null hypothesis?', *Journal of Data Science*, 17(2), pp. 383–390. doi:10.6339/jds.201904_17(2).0008.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p - values. *Royal Society Open Science*, 4(12), p.171085. doi: https://doi.org/10.1098/rsos.171085.

Das D, Das T. (2023). The "P"-value: The primary alphabet of research revisited. International Journal Preventive Medicine, 14(41).

Diaz-Quijano, F.A., Calixto, F.M. and Da Silva, J.M.N. (2020). 'How feasible is it to abandon statistical significance? A reflection based on a short survey,' *BMC Medical Research Methodology*, 20(1). https://doi.org/10.1186/s12874-020-01030-x.

Di Leo, G. and Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, [online] 4(1). doi: https://doi.org/10.1186/s41747-020-0145-y.

Edgeworth, F.Y. (1885). Methods of Statistics, Jubilee volume of the Statist. Soc., E.

Gao, J. (2020). P-values – a chronic conundrum. *BMC Medical Research Methodology*, 20(1). doi: https://doi.org/10.1186/s12874-020-01051-6.

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology*, 45(3), pp.135–140. doi: https://doi.org/10.1053/j.seminhematol.2008.04.003.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical test, p value, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350. DOI: 10.1007/s10654-016-0149-3.

Hurlbert, S. H., Levine, R. A. and Utts, J. (2019). Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires', *The American Statistician*, 73(sup1), pp. 352–357. doi: 10.1080/00031305.2018.1543616.

Kmetz, J. L. (2019). Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values. *The American Statistician*, *73*(sup1), 36–45. https://doi.org/10.1080/00031305.2018.1518271.

Lin, M., Lucas, H.C. and Shmueli, G. (2013). 'research commentary—too big to fail: Large samples and the *p*-value problem', *Information Systems Research*, 24(4), pp. 906–917. doi:10.1287/isre.2013.0480.

McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L. (2019). Abandon Statistical Significance. *The American Statistician*, [online] 73(sup1), pp.235–245. doi: https://doi.org/10.1080/00031305.2018.1527253.

Package 'SGPV'. (2022). Available at: https://cran.r-project.org/web/packages/sgpv/sgpv.pdf [Accessed 2 May 2024].

Pericchi, L., Pereira, C., and Perez, M. (2014). Adaptive revised standards for statistical evidence. Proceedings of the National Academy of Sciences of the United States of America. 111. 10.1073/pnas.1322191111.

Robert,D.(2020, June26).Demystifying the p-value. *The Startup*. https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913.

Stahel, W.A., (2021). New relevance and significance measures to replace p-values. PLoS ONE 16(6): e0252991. https://doi.org/10.1371/journal.pone.0252991.

Stewart, T.G. and Blume, J.D. (2019). Second-Generation P-Values, Shrinkage, and Regularized Models. *Frontiers in Ecology and Evolution*, 7. doi: https://doi.org/10.3389/fevo.2019.00486.

Vidgen, B. and Yasseri, T. (2016). 'P-values: Misunderstood and misused', *Frontiers in Physics*, 4. doi:10.3389/fphy.2016.00006.

Wasserstein, R.L., Schirm, A. L. and Lazar, N. A. (2019). 'Moving to a World Beyond "$p < 0.05$"', *The American Statistician*, 73(sup1), pp. 1–19. doi: 10.1080/00031305.2019.1583913.

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, [online] 70(2), pp.129–133. doi: https://doi.org/10.1080/00031305.2016.1154108 .

Zuo, Y., Stewart, T. G., & Blume, J. D. (2022). ProSGPV: An R package for variable selection with second-generation P-values. *F1000Research*, *11*, 58. https://doi.org/10.12688/f1000research.74401.1.