p-Hacking & Exchange of Scientific Information: A Game-Theoretic Approach

Aleš A. Kuběna
Department of Econometrics
Institute of Information Theory and Automation, CAS, Prague
akub@vsup.cz

Abstract

p-Hacking is a bad science practise, when researchers selects statistical hypothesis ex post such that they omits unsignificant results. From the game-theoretic point of view, the exchange of scientific information via publications is a Bayesian game: Each player-experimenter publishes a favourable part of the result of the experiment, but the "denominator" of this result (= number of test and set of negative results) remains his private information. Publishing and citation practice then demotivates researchers to show full and correct results, and favors the p-Hacking-biased results. As a step to the solution, I propose the concept of Compromise Correction. Firstly we adjust the obtained p-values by transforming p $\rightarrow \frac{1}{p}$ Now we model the multiple-testing problem as a cooperative game: For each S - subset of the tests, value of characteristic function is $\nu(S)$= $\underset{i\in S}{\text{Max}}\left(\frac{1}{p_i}\right)$ The idea of compromise correction is to solve the problem of multiple testing, when taking into account the worth of each coalition = each subset of set of experiments. The solution to this problem is the Shapley value:

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n} \frac{1}{(i(i-1))p_i}}$$

$$p_1 \leq p_2 \leq ... \leq p_n$$

The solution of this problem is tractable, keeps the order of values, is robust w.r.t. changes in "tail" (e.g. increasing the number of strongly negative results) and is piecewise-rational. Robustness w.r.t. tail changes is a property that motivates testing and additional questions and publishes results correctly without fear of relativizing the already achieved significant results. Bonferoni's, resp. Šidak's correction, unlike a compromise correction, limits the maximum "safe" number of additional tests

1 Introduction

1.1 Problem: p-hacking & replication crisis

The replication crisis is a crisis of credibility of the published results of scientific experiments. There is a growing suspicion that many of the results reported as statistically correct, were in actual fact the fluctuation of the random component of observing one experiment in one laboratory. We must always take into account the non-zero frequency of non-replicable and randomly emerged results; part of them arises necessarily, on the basis of statistical error or professional misconduct. However, the ratio of studies whose published results had not been repeated even within the maximum imitation of the original laboratory conditions, significantly exceeds the degree of what could be explained by statistical error. Extensive research [1] has succeeded in repeating 25 % of 67 articles; all of them were from the oncology and haematology areas. This *replication crisis* increases research costs into amounts spent on fruitless follow-up clinical trials and, in addition, threatens health as well as confidence in science. A study of a similar type [10] has also shown "resistance" of non-replicable results to the prestige of the journal : " *The reproducibility of published data did not significantly correlate with journal impact factors, the number of publications on the respective target or the number of independent groups that authored the publications.* " [10]

One of the causes is the so-called p-Hacking [15] The idea of p-Hacking is the hypothesis that non-replicable results arise by the researcher hiding some of the experiment's circumstances. Experimenters conceal multiple-testing and publish, without correction, only those conclusions that have been proven to be significant in the experiment. The p-Hacking hypothesis is statistically testable on large data (= p-values from many articles). [2] statistically proved the non-standard behaviour of the p-value curve around the "magical" threshold p=0.05

However, the replication crisis is not reducible to a mere p-value crisis. The problem would not be solved by replacing a p-value by other statistical indicators in scientific outputs. Cherry-picking can be done based on any statistical indicator. For instance, [3] proved limited replicability of effect size of published results. The p-value is advantageous due to its universality and predictable statistical distribution. For negative results, there is an uniform distribution $U(0,1)$ and, for all the results, a mix of uniform and β-distribution [8]

1.2 Cooperative Game Theory: Basic Definitions

The main of this article is to look to *replication crisis* and *p-hacking* from the game-theoretic point of view. So, this chapter contains basic definitions and concepts of the *cooperative game theory.*

Definition: The pair (Ω, v) is a **cooperative game** (in characteristic function form) if Ω is a finite set of players and $\nu : 2^{\Omega} \to \mathbb{R}$ is a characeristic function that assigns to every coalition $S \subseteq \Omega$ an attainable profit $v(S)$ such that $v(\emptyset) = 0$.

A cooperative game is caled

- **aditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) = v(S) + v(T)$.
- **monotone**, if for all $T, R \in 2^\Omega$ with $S \subset T$, $\nu(S) \leq \nu(T)$
- **superaditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) \geq v(S) + v(T)$.
- **subaditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) \leq v(S) + v(T)$.

Superadditivity implies monotonity, but but monotonity does not imply superadditivity. The game class $maxValueGame[]$ examined in the following section is the set of monotones, but generically subaditive games.

Let $\Gamma = \Gamma(\Omega)$ the set of all cooperative games on Ω and by $\Gamma_1 = \Gamma_1(\Omega)$ the subset of all aditive cooperative games on Ω

Definition: A **value** of games is an operator $\Psi : \Gamma \to \Gamma_1$ s.t. $\Psi \circ \Psi = \Psi$

In particular, we define $\Psi_i(v) := \Psi \circ v(\{i\})$. Clearly, $\Psi \circ v$ is uniquelly determined by the numbers $\Psi_i(v)$.

A special case of the value is the **Shapley value**:

Definition (formula): The **Shapley value** is a value ϕ defined by the formula

$$\phi_i \circ v = \sum_{R \supseteq \{i\}} \frac{\Delta_v(R)}{|R|}$$

where $\Delta_R(v) \in \mathbb{R}$ is a **Harsanyi dividend** of the coalition $R \subseteq \Omega$ defined by

$$\Delta_R(v) = \sum_{T \subseteq R} (-1)^{|R|-|T|} v(T)$$

An alternative, but equivalent definition of the Shaplye value is axiomatic. Shapley theorem [13] proves the existence of a unique game-value operator φ assuming it satisfies the following four axioms:

1. **Linearity**: $\varphi(\alpha v + \beta v') = \alpha\varphi(v) + \beta\varphi(v')$ for all $(\Omega, v), (\Omega, v') \in \Gamma$ and $\alpha, \beta \in \mathbb{R}$
2. **Efficiency**: For all games (Ω, v): $\sum_i \varphi_i(v) = v(\Omega)$
3. **Null-player property**: if $i \in \Omega$ is a *null-player*, i.e. $\forall R \subseteq \Omega\ v(R \cup \{i\}) = v(R)$, then $\varphi_i(v) = 0$
4. **Symmetry** (sometimes called *anonymity*): $\varphi(\rho(i))(\rho \cdot v) = \varphi_i(v)$ for every permutation $\rho \in S_\Omega$ (the function $\rho \cdot v$ is defined by $\rho \cdot v(\rho(R)) := v(R)$) for any $R \subseteq \Omega$)

Axioms 1-4 are independent; in [5] and in [12] are examples of values satisfying any 3 of them and not the 4th.

From the geometric point of view, cooperative game is a point of $\mathbb{R}^{2^\Omega}$ and set of all cooperative games Ω is a $2^{|\Omega|} - 1$ dimensional subspace of the vector space $\mathbb{R}^{2^\Omega}$

From the game-theoretic point of view, cooperative game illustrates an economic situation where a coalition profit or cost depends in general on the involved players in a non-aditive way.

Values of games provide a tool how to evaluate the contibutions of the players. In particular, the Shapley value describes a way how to do it in a *fair* way. Linearity means that *fair* value should be linear. In other words, if the same plaers play two games (v_1, v_2) independently, value of every player should be in sum the same as a value of "join" game $v_1 + v_2$ The second axiom is equivalent to the requirement "The maximum coalition will be formed and its profit will be exactly divided". Null player is a player without any benefit of any coalition; null player property means that value of null player should be 0

The axiom of *symmetry* is an expression of equality of all the participating players. This means that the game-value assigned to them is calculated only from their contributions to the coalitions and does not depend on the particular identity of the player.

Once again from the geometric point of view, a value is an operator $\Psi : \Gamma \to \Gamma_1$. First axiom required that Ψ should be linear, i.e. matrix-representable. Harsanyi dividents $(\Delta_R(v))_{R\subseteq\Omega}$ are coefficients in the *unanimity basis* $(u_S)_{\emptyset\neq S\subseteq\Omega}$

$$u_S(R) = \begin{cases} 1 & S \subseteq R \\ 0 & otherwise \end{cases}$$

$(\Delta_R(v))_{R\subseteq\Omega}$ evaluate a net contribution of coallition R to the total profit $v(\Omega)$

Shapley value of u_S is $\frac{1}{|S|}$ for a members of coalllition S and 0 for non-members. The Shapley value divides the net benefit of each coalition, may be negative, among its members.

2 Compromise correction

2.1 Multiple-testing problem

The research design *one experiment - one atomic result* (one null hypothesis OR one estimeted parameter OR one comparison...) is highly inefficient. So, analysis of experimental data usually tests several hypotheses and estimates several parameters. There are many statistical methods for error-controlling of the experiments with multiple testing: Common known Bonfferoni correction p $\to Np$ and similar Šidak's correction $p \to 1-(1-p)^N$ [14], where p is the number of tests. Complex procedures as a [4], [7], [6]. However, there is no incentive mechanism to actually use these procedures, to publish full resut including unsuccessful tests. And simultaneously, each of these procedures rapidly aggravate the score of the basic result when giving more tests. Scientist who only publishes positive results (and conceals negative) is more successful in publishing. And the set of published scientific information is biased. In the sense of [11], the market of scientific informations exchange is poorly designed.

2.2 Compromise correction: idea

The purpose of each correction of p-values is conrolling of probabilities of type I errors (false positives)

The idea of *compromise correction* is to evaluate net contributtion in the sense of cooperative game theory of any test result p_i to the best result $Min_i[p_i]$:

Firstly, we adjust the obtained p-values by transforming them such that the higher value formally means the more convincing results (instead of original ordering *lower value = better results*) $p \to 1/p$. We assume that an individually rational experimenter without interest in credibility, whose primary motivation is to show the outcome as significant as possible, published the most significant result only, without any correction. On the other hand, the rule of multiple testing requires the Bonferoni or another correction. The compromise correction is based on the question of which of the values contributes to the most significant result $Max[\frac{1}{p_i}]$

Let us interpret the problem of the best result as a game over partial results. If the experimenter would only execute a subset S of experiment, his best value vould be $Max_{i\in S}[\frac{1}{p_i}]$. The idea of compromise correction is to solve the problem of multiple testing, when taking into account the worth of each coalition = each subset of set of experiments.The solution to this problem is the *Shapley value.* So we calculate the *Shapley value* for the cooperative $maxValueGame$

$$\text{maxValueGame: } v(S) = \text{Max}\left[X_i = \frac{1}{p_i} : i \in S\right]$$

2.3 Compromise correction: solution

The solution of this problem is tractable: Let $p_1 \leq p_2 \leq ... \leq p_n$, $\boldsymbol{X} = \left(\frac{1}{p_i}\right)_{i=1}^{n}$ Then Shapley value in the coordinate k is

$$\text{Shapley}[\text{maxValueGame}[\mathbf{X}]]_k = \frac{X_k}{k} - \sum_{i>k}^{n} \frac{X_i}{i(i-1)}$$

and compromise correction operator asigns to the k-th best value a corrected value

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n} \frac{1}{i(i-1))p_i}}$$

Proof: Let $X_1 \geq X_2 \geq ... \geq X_n \geq X_{n+1}$

$$maxValueGame[X_1...X_{n+1}] =$$

$$= maxValueGame[X_1 - X_{n+1}, X_2 - X_{n+1}, ...X_n - X_{n+1}, 0] + constantGame[X_{n+1}]$$

where $constantGame[y][S] = y$ for any nonempty coallition S

By linearity, Shapley value of the $maxValueGame[X_1...X_{n+1}]$ = is the sum of Shapley values of two games defined above. For the first game, value 0 is the nullplayer in the sense of Shapley 3rth axiom: $X_i - X_{n+1} \geq 0$ and $Max[S] = Max[S \cup \{0\}]$ for $S \subseteq \{X_1 - X_{n+1}, X_2 - X_{n+1}, ...X_n - X_{n+1}\}$. For the second game, Shapley value is $(\frac{X_{n+1}}{n+1})_{i=1...n}$ according to the symmetry of Shapley value. So

$$\text{Shapley}[\text{maxValueGame}[[(X)_1^n \cup \{X_{n+1}\}]]_k = \frac{X_k}{k} - \sum_{i>k}^{n+1} \frac{X_i}{i(i-1)}$$

and after inverse transform

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n+1} \frac{1}{i(i-1)p_i}}$$

♢

3 Properties of compromise correction

3.1 Mathematical & computational

- Compromise correction **keeps the order of values**. The k-th best value remains the k-th best value after correction

- **Piecewise linearity** of the operator $(X_i)_{i=1}^n \to \text{Shapley}[\text{maxValueGame}[\text{X}]]_i$

  We reassess the data with a more sensitive test, and we assume that only the first result will improve,

  $$p'_1 < p_1 \leq p_2 = p'_2 \leq p_3 ... \leq p_n = p'_n$$

  Then the compromise correction of all the improvements also gives the best result,

  $$\hat{p}'_1 < \hat{p}_1 \leq \hat{p}_2 = \hat{p}'_2 \leq \hat{p}_3 ... \leq \hat{p}_n = \hat{p}'_n$$

  where

  $$\widehat{(p_i)}_{i=1}^n = \text{CompromiseCorrection}[((p_i)_{i=1}^n]$$

3.2 Game-theoretic, Reverse-game-theoretic & motivational

- **Robustness against tail changes**: Let's assume that the experimenter has achieved a significant result, but there is still material left to test additional questions with a low likelihood of becoming significant. When honestly applying Bonferoni's correction, it is preferable not to carry out further analyzes. The reason is the risk of destroying existing and confirmed results

- the Bonferoni correction coefficient increase after each new test. Compromise corrections cause only small bounded fluctuations of the first results. Intuitively, we consider the test and publishing of additional results to be collectively rational. Compromise correction is not in contradiction with individual rationality.

- Copromise correction is a **centralized mechanism**. It needs to collect all the data for the calculation. However, for the above reasons, it motivates the publication of a whole set of p-values better than Bonferoni's correction.

- Bonferoni's correction defines the upper limit of Compromise correction.

4 Limitations and future work

Unfortunately, compromise correction is a decentralized mechanism. From the reverse-game-theoretic point of view, compromise correction is a half-solution only. Compromise correction eliminates the fears and some disadvantages of complying with the rules of good science, but it does not remove the temptation to do so. Compromise correction works as a lifeline for researchers willing to publish correctly, but does not work well as a sticks on those who do not care about the replicability of their published results. However, I hope that in the second plane higher number of published full-set results will help to improve of power of detection of inaccuracies. Construction of decentralized mechanism is the next plan of the research.

The second weak point of compromise correction is its implausible behavior in situation with two or more very similar tests, i.e. with tests with a high a-priori conditional probability $P(T_2 \text{ is significant}|T_1 \text{ is significant})$. For instance Kaplan-Meier test and Cox-regression for the same data. The compromise-correction coefficient for both is implausible close to maximal (but still smaller than Bonferroni!). The usual motivation in this case is not to increase the number of tests, but to find out more of the model parameters.

Further research will focus on the ability to replace the Shapley value with value associated with a pre-defined network structure: Myerson value [9]

5 Acknowledgement

This work was supported by the Excelence project P402/12/G097 DYME Dynamic Models in Economics of GAČR

References

[1] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531, 2012.

[2] S. B. Bruns and J. P. Ioannidis. P-curve and p-hacking in observational research. *PLoS One*, 11(2):e0149144, 2016.

[3] O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[4] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, pages 111–139, 2002.

[5] R. P. Gilles. *The cooperative game theory of networks and hierarchies*, volume 44. Springer Science & Business Media, 2010.

[6] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[7] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[8] L. R. Jager and J. T. Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, 2013.

[9] R. B. Myerson. On the value of game theory in social science. *Rationality and Society*, 4(1):62–73, 1992.

[10] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712, 2011.

[11] A. E. Roth. *Who Gets What–and why*. Brilliance Audio, 2015.

[12] D. Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on applied mathematics*, 17(6):1163–1170, 1969.

[13] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[14] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[15] U. Simonsohn, L. D. Nelson, and J. P. Simmons. P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534, 2014.