Evidence of p-value fiddling using a rapid, high-volume systematic method

Kroeger CM[1], Brown AW[2], Allison DB[2]

Department of Kinesiology and Nutrition, University of Illinois at Chicago, IL, USA[1]

Nutrition Obesity Research Center, University of Alabama at Birmingham, AL, USA[2]

Abstract

Background P-value fiddling, also called 'researcher degrees of freedom' can entail manipulating statistical analyses until the p-value is below the nominal α level. Because not all studies publicly register a pre-specified analysis plan, the extent of p-value fiddling in scientific literature is unknown. Manually extracting the large number of p-values needed for hypothesis testing in this domain can be unwieldy.

Objective We utilized PubMed search infrastructure to implement a rapid, high-volume method to gain insight about evidence of p-value fiddling within scientific literature.

Design We searched PubMed for: 1) p-values within the ranges, 0.03*, 0.04*, 0.05*, and 0.06*, 2) what we call common statistical analyses (e.g. t-tests, ANOVAs, ANCOVAs), and 3) analyses we call atypical (e.g. Wilcoxon-rank, data transformations, outlier removal). We excluded abstracts containing more than one p-value range or analysis type. We tested for an association between p-value category and analysis choice via contingency table analysis.

Results For atypical analyses, the observed frequency of 0.04* p-values exceeded the expected values by 3.5%, whereas those in the 0.05* range occurred 21.2% less often than expected (p=0.02). This relationship was inverted for common analyses.

Conclusion P-values in the 0.04* range are overrepresented and those in the 0.05* range are underrepresented when atypical statistical methods are used. This suggests that when nearly statistically significant results are obtained with a typical test, investigators may preferentially switch to an atypical test and report that result if it is statistically significant. Using simple search tools can provide quick insights into meta-research questions. Our next step will be to refine this method to better measure studies' choice of statistical approach while retaining speed and efficiency.

Background & Rationale

P-value fiddling, also called 'p-hacking' and 'researcher degrees of freedom' can entail reanalyzing datasets using various tests until a p-value of interest is below the nominal α level, especially without disclosing that one is doing so [1-3]. Investigators are encouraged to publically register a pre-specified statistical analysis plan to help decrease p-value fiddling and the chances of obtaining false positive results, however many investigators do not do this [4,5].

When nearly statistically significant results are obtained with a typical test, investigators may preferentially switch to an atypical test and report that result if it is statistically significant. Evaluating the distribution of p-values among common statistical analyses (e.g., t-tests, ANOVAs, and ANCOVAs) versus atypical tests (e.g., Wilcoxon-rank, Kruskal-wallis, data transformations, and outlier removal) may provide valuable insight about evidence of p-value fiddling within scientific literature. Because many investigators do not publically register a pre-specified analysis plan, the extent that p-value fiddling occurs within scientific literature is unknown.

Previous work has shown that specific statistical tests can be used to detect p-value fiddling within scientific literature [6]. However, manually extracting the large number of p-values needed for hypothesis testing in this domain can be unwieldy. Whether simple search tools can be used to increase the magnitude, speed, and efficiency of data collection, while still providing meaningful information about evidence of p-value fiddling within scientific literature, has yet to be shown.

Objective

To develop a high-speed automated method to test the magnitude with which p-value distributions differ between common statistical analyses and atypical statistical analysis, to gain insight about evidence of p-value fiddling within scientific literature, under the assumption that the use of at least some atypical statistical analyses were not pre-specified. We therefore utilized PubMed search infrastructure to implement a rapid, high-volume method to evaluate the scientific literature.

Methods

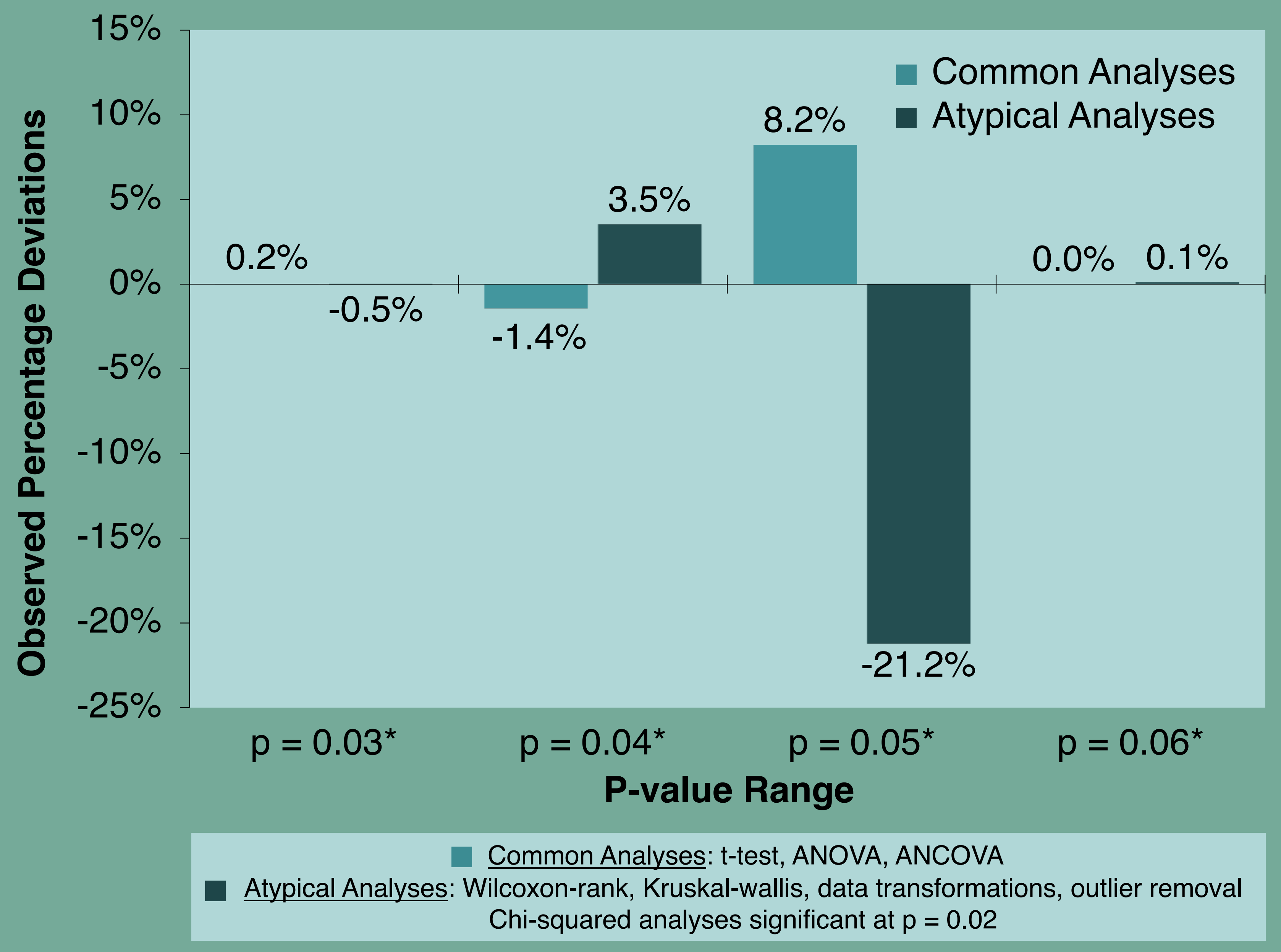
We searched over 24 million abstracts in PubMed for: 1) p-values reported within four ranges, 0.030-0.039 (0.03*), 0.040-0.049 (0.04*), 0.051-0.059 (0.05*), and 0.060-0.069 (0.06*), 2) common statistical analyses and 3) atypical analyses. We excluded abstracts containing more than one p-value range or analysis type, in order to isolate each. We then searched each p-value range in combination with each analysis group to determine the prevalence of combinations within PubMed. We systematically analyzed 476,613 abstracts that fit our inclusion criteria, and 8,947 abstract combinations qualified for analysis. We tested for an association between p-value category and analysis using a chi-squared test.

Results

Table 1. Search Query in PubMed

Search	PubMed Query		Items	Time
#22	#14 AND #12	Atypical + 0.06*	181	11:18
#21	#14 AND #11	Atypical + 0.05*	103	11:17
#20	#14 AND #10	Atypical + 0.04*	1011	11:17
#19	#14 AND #9	Atypical + 0.03*	1209	11:17
#18	#13 AND #12	Common + 0.06*	465	11:16
#17	#13 AND #11	Common + 0.05*	364	11:16
#16	#13 AND #10	Common + 0.04*	2480	11:16
#15	#13 AND #9	Common + 0.03*	3134	11:15
#14	#8 NOT #7		162910	11:11
#13	#7 NOT #8		83881	11:11
#12	#6 NOT (#1 OR #3 OR #5)		17059	11:10
#11	#5 NOT (#1 OR #3 OR #6)		10323	11:08
#10	#3 NOT (#1 OR #5 OR #6)		90871	11:07
#9	#1 NOT (#3 OR #5 OR #6)		111569	11:06
#8	Search nonparametric [tiab] OR non-parametric [tiab] OR wlicoxon-rank [tiab] OR "wilcoxon rank" [tiab] OR kruskal-wallis [tiab] OR "kruskal wallis" [tiab] OR transformation [tiab] OR (outlier* [tiab] AND remov* [tiab])		166384	11:04
#7	Search t-test [tiab] OR anova [tiab] OR ancova [tiab] OR "mixed model" [tiab]		87355	11:02
#6	Search p=.06* [tiab] OR p=0.06* [tiab]		25343	11:01
#5	Search (p 05,05[tiab] OR p 050[tiab] OR p 0500[tiab] OR p 05006[tiab] OR p 0501[tiab] OR p 0501a[tiab] OR p 0502[tiab] OR p 0505[tiab] OR p 0507[tiab] OR p 051[tiab] OR p 0511[tiab] OR p 0513[tiab] OR p 0514[tiab] OR p 0515[tiab] OR p 0518[tiab] ORp 0519[tiab] OR p 052[tiab] OR p 0529[tiab] OR p 053[tiab] OR p 0537[tiab] OR p 054[tiab] OR p 0542[tiab] OR p 055[tiab] OR p 0557[tiab] OR p 056[tiab] OR p 0566[tiab] OR p 057[tiab] OR p 0571[tiab] OR p 0573[tiab] OR p 058[tiab] OR p 0583[tiab] OR p 0585[tiab] OR p 0588[tiab] OR p 059[tiab] OR p 0591[tiab] OR p 0594[tiab] OR p 0595[tiab] OR p 0597[tiab]) OR (p 0 05,15[tiab] OR p 0 05,863[tiab] OR p 0 05,95[tiab] OR p 0 050[tiab] OR p 0 0500[tiab] OR p 0 0501[tiab] OR p 0 0502[tiab] OR p 0 0503[tiab] OR p 0 0504[tiab] OR p 0 0505[tiab] OR p 0 0505385[tiab] OR p 0 0506[tiab] OR p 0 0507[tiab] OR p 0 0508[tiab] OR p 0 0509[tiab] OR p 0 051[tiab] OR p 0 0510[tiab] OR p 0 0511[tiab] OR p 0 0512[tiab] OR p 0 0513[tiab] OR p 0 0514[tiab] OR p 0 0515[tiab] OR p 0 0516[tiab] OR p 0 0517[tiab] OR p 0 0518[tiab] OR p 0 0519[tiab] OR p 0 052[tiab] OR p 0 0520[tiab] OR p 0 0522[tiab] OR p 0 0523[tiab] OR p 0 0524[tiab] OR p 0 0525[tiab] OR p 0 052574[tiab] OR p 0 0526[tiab] OR p 0 0527[tiab] OR p 0 0529[tiab] OR p 0 053[tiab] OR p 0 0530[tiab] OR p 0 0531[tiab] OR p 0 0532[tiab] OR p 0 0533[tiab] OR p 0 0534[tiab] OR p 0 0535[tiab] OR p 0 0536[tiab] OR p 0 0537[tiab] OR p 0 0539[tiab] OR p 0 054[tiab] OR p 0 0540[tiab] OR p 0 0541[tiab] OR p 0 0542[tiab] OR p 0 0543[tiab] OR p 0 0544[tiab] OR p 0 0545[tiab] OR p 0 0546[tiab] OR p 0 0547[tiab] OR p 0 0548[tiab] OR p 0 0549[tiab] OR p 0 055[tiab] OR p 0 0550[tiab] OR p 0 0551[tiab] OR p 0 0552[tiab] OR p 0 0553[tiab] OR p 0 0554[tiab] OR p 0 05548[tiab] OR p 0 0556[tiab] OR p 0 0557[tiab] OR p 0 0558[tiab] OR p 0 0559[tiab] OR p 0 056[tiab] OR p 0 0560[tiab] OR p 0 0561[tiab] OR p 0 0562[tiab] OR p 0 0563[tiab] OR p 0 0564[tiab] OR p 0 0566[tiab] OR p 0 0567[tiab] OR p 0 0568[tiab] OR p 0 0569[tiab] OR p 0 057[tiab] OR p 0 0570[tiab] OR p 0 0571[tiab] OR p 0 0572[tiab] OR p 0 0573[tiab] OR p 0 0574[tiab] OR p 0 0575[tiab] OR p 0 0576[tiab] OR p 0 05760[tiab] OR p 0 0577[tiab] OR p 0 0578[tiab] OR p 0 0579[tiab] OR p 0 058[tiab] OR p 0 0580[tiab] OR p 0 0581[tiab] OR p 0 0582[tiab] OR p 0 0583[tiab] OR p 0 0584[tiab] OR p 0 05846[tiab] OR p 0 0585[tiab] OR p 0 0586[tiab] OR p 0 0587[tiab] OR p 0 0588[tiab] OR p 0 0589[tiab] OR p 0 05891[tiab] OR p 0 059[tiab] OR p 0 0590[tiab] OR p 0 0591[tiab] OR p 0 0592[tiab] OR p 0 0593[tiab] OR p 0 0594[tiab] OR p 0 0595[tiab] OR p 0 0596[tiab] OR p 0 0597[tiab] OR p 0 0598[tiab] OR p 0 0599[tiab] OR p 0 05a[tiab] OR p 0 05and[tiab] OR p 0 05for[tiab] OR p 0 05u[tiab] OR p 0 05vs[tiab] OR p 0 05vsho[tiab])		16409	10:59
#3	Search p=.04* [tiab] OR p=0.04* [tiab]		130032	10:55
#1	Search p=.03* [tiab] OR p=0.03* [tiab]		151564	10:53

Figure 1. Deviation of Observed P-value Distributions from Expected Distributions within PubMed Scientific Abstracts for Common versus Atypical Statistical Analyses



🟦 Common Analyses: t-test, ANOVA, ANCOVA

⬛ Atypical Analyses: Wilcoxon-rank, Kruskal-wallis, data transformations, outlier removal

Chi-squared analyses significant at p = 0.02

Conclusions

The evidence found using this method is consistent with our hypothesis that p-values in the 0.04* range are overrepresented in the literature and p-values in the 0.05* range are underrepresented when atypical statistical methods are used. This indicates potential evidence of p-value fiddling among studies reporting results from atypical tests. Additionally, we have found that using simple search tools can provide quick insights into meta-research questions. Our next steps will be to refine the method to better measure studies' choice of statistical approach and obtained p-values, while still retaining most of the speed and efficiency of the fully automated method.

References

[1] Motulsky HJ. Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Arch Pharmacol* (2014) 387:1017-1023. **[2]** Simonsohn U, et al. P-curve: a key to the file drawer. *J Exp Psychol Gen* (2014) 143 (2):534-47. **[3]** Masicampo EJ, et al. A peculiar prevalence of p values just below .05. *Q J Exp Psychol (Hove)* (2012) 65(11):2271-9. **[4]** Simmons JP, et al. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* (2011) 22(11):1359-1366.

[5] Murayama K, et al. Research practices that can prevent an inflation of false-positive rates. *Pers Soc Psychol Rev* (2013) **[6]** Gadbury GL, et al. Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. *Plos One* (2012) 7(10):e46363.