Hassler, Uwe; Pohle, Marc-Oliver

Unlucky Number 13? Manipulating Evidence Subject to Snooping

International Statistical Review

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Hassler, Uwe; Pohle, Marc-Oliver (2022) : Unlucky Number 13? Manipulating Evidence Subject to Snooping, International Statistical Review, ISSN 1751-5823, Wiley, Hoboken, NJ, Vol. 90, Iss. 2, pp. 397-410, https://doi.org/10.1111/insr.12488

International Statistical Review

International Statistical Review (2022), **90**, 2, 397–410 doi: 10.1111/insr.12488

Unlucky Number 13? Manipulating Evidence Subject to Snooping

Uwe Hassler and Marc-Oliver Pohle

Goethe University Frankfurt, RuW Building, Theodor-W.-Adorno-Platz 4, Frankfurt, 60323, Germany
Marc-Oliver Pohle, Goethe University Frankfurt, RuW Building, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany. Email: pohle@econ.uni-frankfurt.de

Summary

Questionable research practices have generated considerable recent interest throughout and beyond the scientific community. We subsume such practices involving secret data snooping that influences subsequent statistical inference under the term MESSing (manipulating evidence subject to snooping) and discuss, illustrate and quantify the possibly dramatic effects of several forms of MESSing using an empirical and a simple theoretical example. The empirical example uses numbers from the most popular German lottery, which seem to suggest that 13 is an unlucky number.

Key words: data snooping; p-hacking; p-values; research transparency.

1 Introduction

We propose the generic term MESSing (manipulating evidence subject to snooping) to subsume practices that involve conducting statistical inference after data analysis of some form has already been carried out and influenced the researchers' decisions, but is not acknowledged. In recent years, a number of such questionable research practices related to statistical inference situated somewhere in the grey zone between exploratory data analysis and fraud have attracted the attention of the academic literature in many fields such as medicine, psychology or economics (see, e.g. Ioannidis, 2005; Simmons *et al*., 2011; and Brodeur *et al*., 2016). Often these practices seem innocent, are not used with bad intentions and are deeply rooted in the research culture. But they invalidate inference and may consequently lead to wrong results and distorted literatures. Further, due to the lack of transparency related to these practices, their extent and exact consequences are very hard to assess. Their detrimental effects like, for example, impeding the replicability of research, slowing down scientific progress or damaging the credibility of science have been discussed and some steps have been taken to curb them in recent years (see, e.g. Wasserstein *et al*., 2019 or Christensen *et al*., 2019). Still, they seem to be in widespread use. Our paper aims to contribute to the understanding of the issue and the problems arising from it. We illustrate our overarching definition of MESSing with two cases, which demonstrate the possibly dramatic effects of seemingly innocent or small manipulations on the validity of statistical inference and showcase different forms of MESSing. We also review and connect the literature on questionable research practices of that kind. The first example studies German lottery numbers, where at first glance surprisingly the number 13 is drawn significantly

less often than expected under uniformity. Second, we discuss as a simple theoretical example different strategies when testing for normality, where MESSing may consist of maximising or just as well of minimising evidence. We hence stress the case of MESSing where evidence is driven to the extreme instead of just jumping over certain significance thresholds.

The next section addresses how the 'puzzle' of an overly significant unlucky number 13 arises, how it can be solved and how this relates to MESSing. Section 3 connects to previous literature and goes into some details on MESSing. Section 4 contains the simple theoretical example that illustrates and allows to quantify the effects of MESSing in different directions. Some conclusions are offered in the final section.

2 Testing for Unlucky 13

The most popular lottery in Germany is 'Lotto 6 out of 49'. The 6 winning numbers of one game are determined by drawing 6 balls without replacement from a pool of 49 balls identified by means of the first 49 natural numbers. The first game took place on 9 October 1955, and the first ball ever drawn carried the number 13. Until 29 November 2019, a total of $N = 4337$ games had been played with $n = 6 \cdot 4337 = 26022$ balls being drawn. Figure 1 displays the absolute frequencies for each number 1 through 49.[1] What does strike you at first glance? The number 13 stands out with the least favourable odds. This may come as no surprise to people that consider 13 to be an 'unlucky number'. As a result of fear of the number 13 (clinically: triskaidekaphobia), there is no row 13 in many planes or no floor 13 in many tall buildings and many people avoid Friday the 13th for marriage. And indeed, the number 13 was drawn only 471 times in the German lottery, while (roughly) 531 cases would have to be expected under equal probability of all 49 numbers given 26022 draws (broken line in Figure 1). If a PhD student presents such descriptive evidence to her or his supervisor, the supervisor might ask: Is the deviation significant? At which level? How can we test properly? And how can we avoid the so-called Texas sharpshooter fallacy (where the shooter paints the target after firing a shot)?

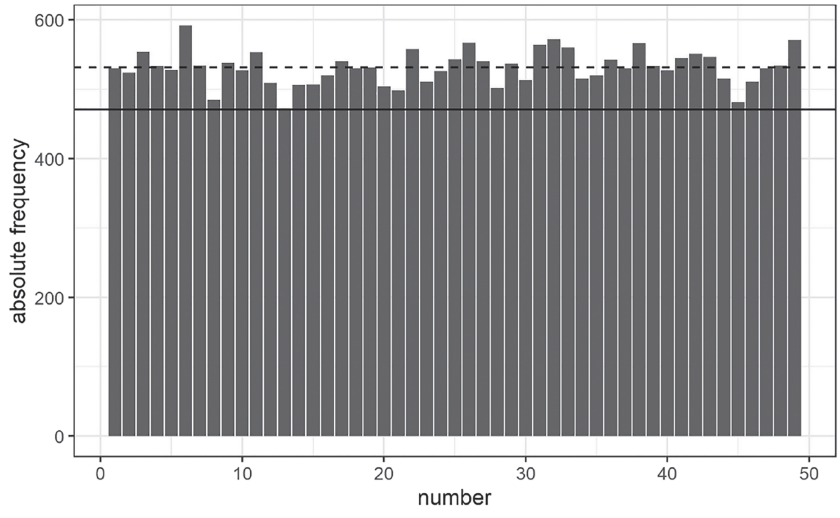

FIGURE 1. *Frequency distribution of $N = 4337$ games, that is, from $n = 6 \cdot N = 26022$ numbers (German Lotto 6 out of 49); broken line: expected frequency under equal probability, solid line: frequency of number 13*

To execute a test of uniformity, we are interested in the counts of each of the 49 numbers from a sample of size $n = N \cdot 6$. Let the counts of these numbers be denoted by S_m, $m = 1, 2, \ldots, 49$. Let p_m be the probability of getting the ball numbered by m when drawing a Lotto number. First, we are interested in testing the null hypothesis

$$H_0 : p_m = \frac{1}{49} \text{ for one specific } m \in \{1, 2, \ldots, 49\}.$$

Consider the classical binomial test statistic constructed under the i.i.d. assumption,

$$Z_m^{iid} := \frac{S_m - \frac{n}{49}}{\sigma_{iid}} \text{ with } \sigma_{iid}^2 := \frac{n(49-1)}{49^2}. \tag{1}$$

Due to the dependence between the six draws of one game, it does not follow a standard normal law asymptotically. However, as we show in the appendix, only the variance decreases due to the negative dependence. Define (as special case of equation (A1) in the appendix) the variance

$$\sigma_{lot}^2 := \frac{49-6}{49-1}\, \sigma_{iid}^2, \tag{2}$$

and limiting standard normality is retained under H_0 (see (A2) in the appendix for a discussion of the general case beyond the German lottery):

$$Z_m^{lot} := \frac{S_m - \frac{n}{49}}{\sigma_{lot}} = \sqrt{\frac{49-1}{49-6}}\, Z_m^{iid} \xrightarrow{d} \mathcal{Z}, \tag{3}$$

where $\mathcal{Z}$ follows a standard normal distribution, $\mathcal{Z} \sim \mathcal{N}(0,1)$, and '$\xrightarrow{d}$' denotes convergence in distribution as the sample size n diverges.

Now we are equipped to return to the data with a sample of size $n = 6 \cdot 4337 = 26022$. We wish to test $p_{13} = 1/49$ against the one-sided alternative:

$$H_0 : p_{13} = \frac{1}{49} \text{ vs. } H_1 : p_{13} < \frac{1}{49}.$$

The test statistic accounting for the implied dependence of the German lottery due to drawing without replacement from equation (3) results in $Z_{13}^{lot} = -2.7822$ with the highly significant (one-sided) p-value of 0.00270 relying on the normal approximation. Is the German lottery flawed? Is 13 truely an unlucky number? What is going wrong? What causes this test result is of course MESSing. This nonsensical significance arises because we first looked at the data in Figure 1, observed the remarkable deviation of $S_{13} = 471$ from $26022/49 = 531.06$ and then tested for the specific hypothesis $p_{13} = 1/49$. A real MESSy might try to tell a convincing story why H_1: $p_{13} < 1/49$ is a plausible alternative a priori, which indeed found highly significant support by the data. Such a MESS had been blamed already by Wallis (1942, p. 229): 'An investigator who after inspecting the data decides what to test or how to make the test can, by virtue of the fact that any sample has unique characteristics, disprove any hypothesis.'

The number 13 was picked for testing because $m = 13$ leads to the strongest left-sided violation of the null in favour of $p_m < 1/49$. We are able to quantify the effect of this MESS assuming that the Lotto numbers follow a uniform distribution. What we did amounts to testing with

$\min_{m=1,\,\ldots,\,49} S_m$, and not surprisingly, the minimum deviates significantly from the overall mean. Let

$$Z_{\min} := \frac{\min\limits_{m\,=\,1,\,\ldots,\,49} S_m \;-\; \dfrac{n}{49}}{\sigma_{lot}}.$$

We denote the limit in distribution of $Z_{\min}$ for $n \to \infty$ as $\mathcal{Z}_{\min}$. The density of $\mathcal{Z}_{\min}$ is depicted in Figure 2 alongside a standard normal density, that is, the density of the asymptotic distribution of Z_m^{lot} under the null. The density of $\mathcal{Z}_{\min}$ was obtained by simulations; details are described in the appendix.[2] Due to MESSing, the distribution dramatically changes its shape, leading to very small p-values. From $\mathcal{Z}_{\min}$, we can also calculate the size distortions: the rejection probabilities under the null hypothesis, that is, the type I errors for the three significance levels 0.01, 0.05 and 0.1 are (with z_α denoting the α-quantile of $\mathcal{Z} \sim \mathcal{N}(0,1)$):

$$P(\mathrm{Z}_{\min} \le z_{0.01}) = 0.3879,\; P(\mathrm{Z}_{\min} \le z_{0.05}) = 0.9376 \text{and}\;\; P(\mathrm{Z}_{\min} \le z_{0.1}) = 0.9984.$$

Of course we have a method that is robust to MESSing in our example. A proper way to test the null hypothesis of uniformity has to take into account *each* number $m \in \{1, \ldots, 49\}$, which amounts to the goodness-of-fit test by Pearson (1900) for the joint hypothesis:

$$H_0 : p_1 = p_2 = \ldots = p_{49} = \frac{1}{49}. \tag{4}$$

Again the dependence of the six draws within one game due to drawing without replacement invalidates the classical approach: the test statistic

$$\chi^2_{iid} := \sum_{m=1}^{49} \frac{\left(S_m \;-\; \frac{n}{49}\right)^2}{\dfrac{n}{49}} \tag{5}$$

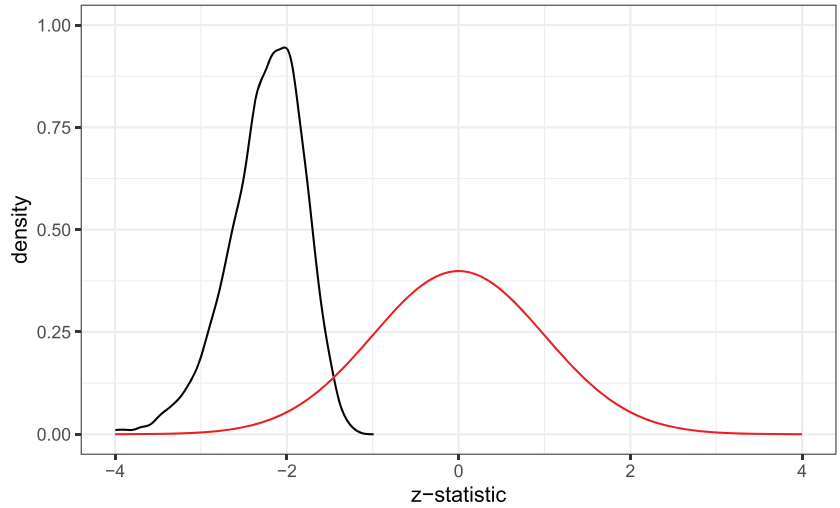

FIGURE 2. *Density of $\mathcal{Z}_{\min}$ (obtained by simulations) compared with a standard normal density*

does not converge in distribution to a $\chi^2(48)$ random variable. And again, only a scaling factor (known from equation (2)) is required to recover the limiting chi-squared distribution under the null:

$$\chi^2_{lot} := \frac{49 - 1}{49 - 6}\chi^2_{iid} \xrightarrow{d} \chi^2(48). \tag{6}$$

This limit arises as a special case of more general results by Joe (1993, p. 183) for Pearson's tests for uniformity of k-tuples, $k \in \{1, 2, \ldots, K\}$ in the lottery K out of M; see also Genest *et al*. (2002) or the earlier closely related results from the survey sampling literature by Rao & Scott (1981). The data behind Figure 1 provide $\chi^2_{iid} = 55.10$ and $\chi^2_{lot} = 61.51$. The p-value of χ^2_{lot} when comparing with the $\chi^2(48)$ distribution is 9.11%. The p-value is of course much higher than when testing against $p_{13} < 1/49$, and the German lottery data do not violate the uniform distribution hypothesis (4) at a 9% significance level.

The firm operating the lottery may not be happy with a p-value of 9.11%, which is below the 10% hurdle that many researchers maintain for a semblance of significance. To leave no doubt about the uniform distribution, the Lotto operator might wish to produce weaker evidence, that is a larger p-value. This can easily be achieved by changing the sample. Notwithstanding the name '6 out of 49', in 3615 out of 4337 games, an additional number was drawn, which changed the price money,[3] amounting in fact to 7 numbers drawn from 49 without replacement. The additional numbers form a new sample of size $n = 3615$, because in $N = 3615$ games just one ball is drawn. The frequencies of this new sample are depicted in Figure 3. Note that the observations in this sample of additional numbers are independent. Hence, the null hypothesis (4) can be tested with Pearson's conventional statistic. It yields $\chi^2_{iid} = 48.64$ with a p-value of 44.7% being beyond any reasonable significance level. If the rationale is not only to produce a p-value above 5% but really to minimise the evidence against uniformity, that is, to be as far away from significance as possible in terms of the p-value, the lottery firm may come up with some arguments to restrict the attention to the sample of additional numbers only. This might be seemingly justified by the fact that this sample of size $n = 3615$ is independent such that the conventional χ^2_{iid} from (5) may be used for testing.

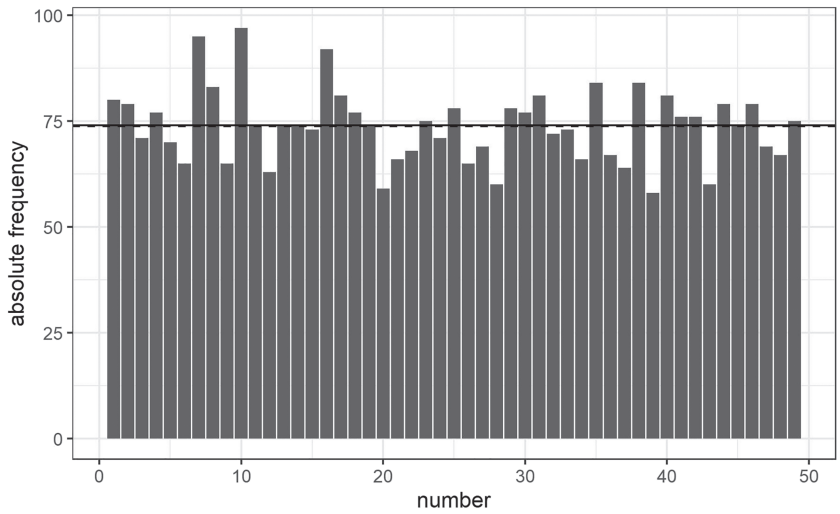

FIGURE 3. *Frequency distribution of the additional numbers from 3615 games*

Let us summarise the three outcomes of this section: (1) moderate or weak significance of χ^2_{lot} from the dependent sample behind Figure 1 (p-value of 9.11%), (2) very high significance (p-value of 0.27%) of Z^{lot}_{13} computed from this sample and (3) clear insignificance by any convention of χ^2_{iid} from the sample of additional numbers (p-value of 44.70%). This illustrates the effects of MESSing and how MESSies may proceed in practice. In (1), we refrain from cheating but still obtain significance at the 10% level. In (2), we manipulate the hypothesis or choice of test subject to snooping to maximise the evidence against the null. In (3), we are not happy with merely jumping over some (in)significance threshold, but wish to drive evidence in terms of p-values to the extremes, that is, to minimise the evidence against the null. Here, MESSing comes up with some flimsy excuse why to move to the new sample of additional numbers only.

3 Manipulating Evidence Subject to Snooping

By coining the term, we want to stress that it is crucial to distinguish MESSing from exploratory data analysis and data mining on the one hand and outright fraud on the other hand. The term 'data mining' has undergone a considerable change in meaning. Lovell (1983) used it to describe the process of 'experimentation' until t-statistics turn significant, sometimes called data grubbing or dredging or fishing to defame the applied practice of others; see also White (2000, p. 1098): 'Although data mining has recently acquired positive connotations as a means of extracting valuable relationships from masses of data, the negative connotations arising from the ease with which naive practitioners may mistake the spurious for the substantive are more familiar to econometricians and statisticians.' However, nowadays, data mining receives a lot of attention as a smart toolkit for computational data analysis intersecting with machine learning and engineering techniques like artificial intelligence in order to unveil hidden association or patterns in big data sets; see, for example, Hastie *et al*. (2009) for an appreciation. Thus, the term data mining is today often interpreted as a form of exploratory data analysis for big data (see Hand, 1998). Exploratory data analysis is crucial for virtually any statistical analysis, and in other applications, the data set is often not large enough to split the sample for exploratory analysis and inference; thus, data snooping is 'endemic' not only in time series analysis (see White, 2000). MESSing is the dark side of the perfectly sensible and necessary practices of exploratory data analysis, data mining or snooping; it applies when they may become harmful, namely, when they are not acknowledged. On the other hand, even though not acknowledging data snooping is clearly a manipulation, it should be distinguished from outright fraud like, for example, fabrication of data.

We now discuss some forms of MESSing and related issues in detail: one way to MESS is HARKing (Hypothesizing After the Results are Known) defined by Kerr (1998, p. 197) as '[…] presenting post hoc hypotheses in a research report as if they were, in fact, a priori hypotheses'. Generally, statistical tests are invalidated when one postulates hypotheses or test statistics subject to data snooping and uses the same data to test them, because 'agreement between a sample and a hypothesis based on that sample is purely tautological and proves nothing but accuracy in reading and restating the data of the sample'; see Wallis (1942, p. 229). The traditional approach of statistical testing had been designed for what Hand (1998, p. 112) called primary data analysis: '[…] the data are collected with a particular question or set of questions in mind.' This is the reason why subfields like sampling theory and experimental or survey design are central to statistical theory and practice. Hand (1998) distinguished secondary data analysis defined as '[…] the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners'. Kerr (1998) and Rubin (2017) demonstrated that data mining, or secondary data analysis, invalidates

hypothesis testing: a hypothesis that has been postulated only after explorative data inspection must not be tested with the same data. This is why Hollenbeck & Wright (2017) distinguished between THARKing (transparently HARKing) and SHARKing (secretly HARKing). They classified SHARKing as an unethical practice. Clearly, not all researchers share this view, and SHARKing may be a widespread 'questionable research practice' as investigated by John *et al*. (2012) in an anonymous survey of more than 2000 academic psychologists. John *et al*. (2012, table 1) observed that 35% affirmed of 'reporting an unexpected finding as having been predicted from the start', which is in the spirit of SHARKing of course.

A further popular form of MESSing is what has been called *p*-hacking recently; see Simonsohn *et al*. (2014, p. 534): dredging the data until the *p*-value is small enough to reject; see also Simmons *et al*. (2011). That way one may produce 'spectacular results' and catch attention of a wider public. But also in the smaller scientific community, there are strong incentives to produce 'false positive' results, due to the so-called publication bias. Sterling (1959, p. 30) already stated that: 'There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs–an "error of the first kind"–and is published'; see also Sterling *et al*. (1995). In times where researchers are under increasing pressure to publish successfully, HARKing and *p*-hacking may be all the more tempting because 'negative results' (nonrejection of hypotheses) are hard to publish. A number of studies have tried to quantify the amount of *p*-hacking in a certain discipline or literature by collecting *p*-values or values of test statistics from a large number of papers and analysing their empirical distributions: Gerber & Malhotra (2008a) and Gerber & Malhotra (2008b) observe large jumps after the classical significance thresholds in the empirical distribution of *Z*-statistics from papers from political science and sociology and interpret this as evidence for the prevalence of *p*-hacking in these disciplines. Brodeur *et al*. (2016) and Brodeur *et al*. (2020) consider papers from economics and find a trough in the distribution of *Z*-statistics corresponding to *p*-values between 10% and 25%, interpret this as evidence for *p*-hacking and try to draw conclusions regarding the amount of *p*-hacking in economics. Focussing on a single study or a small number of studies instead of on an aggregate, replication is a further, but very costly, way to analyse the reliability of studies and to possibly find evidence for the use of questionable research practices (see, e.g. Christensen *et al*., 2019, chapter 9).

Sometimes, however, the incentives may be the other way around as well: a researcher may be happy not to reject a null hypothesis. Consider specification testing of certain assumptions behind a model we wish to apply. If, for example, we want to perform a simple analysis of variance (ANOVA), the underlying assumptions are normality of the data and variance homogeneity. A conscientious statistician would check these assumptions before applying the ANOVA *F*-test, and he or she might be tempted to weaken evidence against the underlying assumptions to jump over the chosen significance level. Such a behaviour has been called reverse *p*-hacking by Chuard *et al*. (2019). Reverse *p*-hacking might also be observed in research influenced by industries, which are interested in weakening evidence of negative effects of their products on health, for example, of cigarette smoking on lung cancer (see, e.g. White & Bero, 2010).

There may be several reasons why researchers do not only want to reach *p*-values just below (or above) classical significance levels as in the case of (reverse-)*p*-hacking but to really minimise (or maximise) *p*-values. For example, smaller *p*-values often are perceived as lending more credibility and importance to results without even looking at the respective effect sizes (see, e.g. Wasserstein & Lazar, 2016) or being far away from significance levels could lead to being above suspicion of *p*-hacking.

To avoid the negative effects of SHARKing or *p*-hacking or of MESSing in general, Wasserstein & Lazar (2016, p. 132) demanded more transparency from scientific authors: 'Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.' Nelson *et al*. (2018) also put forward disclosure as one of the two main remedies against *p*-hacking. As the other they mention preregistration of research plans, where researchers specify the main statistical analyses they intend to do before they collect their data. This is certainly one of the most effective measures against questionable research practices. For experimental studies, preregistration is widely used in medicine and has recently become popular in psychology and economics as well (see, e.g. Christensen *et al*., 2019, chapter 6). However, preregistration is not really credible for analyses of existing data and thus unfortunately can hardly serve as a solution there.

On top of requirements for authors, Simmons *et al*. (2011, p. 1363) added guidelines for reviewers and recommended not to push the authors to provide highly significant results. Similarly, Sterling *et al*. (1995, p. 111) encouraged journal editors to accept or reject empirical studies irrespective of their outcomes and to make decision rather in light of the importance of the research question and the adequacy of the employed methods and data. In 2015, the editors of *Basic and Applied Social Psychology* (BASP) went one step further and virtually banned '*p*-values, *t*-values, *F*-values, statements about "significant" differences or lack thereof, and so on' from this journal; see the Editorial by Trafimow & Marks (2015, p. 1). Unfortunately, a ban of signifiance rituals opens different routes to questionable research practices. In particular, Fricker Jr. *et al*. (2019, p. 374) found when assessing all papers published in BASP in 2016 '[…] multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered'.

Sometimes other statistical methods are suggested to replace or complement *p*-values as a means of mitigating questionable research practices. Nelson *et al*. (2018) criticise this and provide a short summary and references on this debate. In particular, Simonsohn (2014) demonstrates that Bayesian methods are no remedy against MESSing. A natural replacement or complement for *p*-values in general are confidence intervals as they provide more information, especially an assessment of estimation uncertainty (see, e.g. Romer, 2020 or Coulson *et al*., 2010, and the literature review therein). Note that confidence intervals may of course also be subject to manipulation and are included in our definition of MESSing presented at the beginning of the article, which is not about testing specifically but about statistical inference in general. For example, a researcher may want to present very precise estimation results and thus secretly choose the specification of a regression model, which leads to the narrowest confidence intervals on the parameter of interest. When it comes to mitigating questionable research practices, confidence intervals may nevertheless help in that they could contribute to moving away from the narrow focus on the dichotomous testing decision and thus de-emphasise statistical significance and reduce the incentives for MESSing (see, e.g. Imbens, 2021). However, as this focus on testing and significance thresholds is deeply rooted in research culture and statistical education in many disciplines, often researchers will probably just use the confidence interval to check if the hypothesised parameter value falls inside the confidence interval, in which case it would of course make no difference if *p*-values from two-sided tests or confidence intervals were reported.

We close this section by mentioning two related issues, post model selection inference and multiple testing, and stressing the importance of statistical education in mitigating questionable research practices. In practice, the model behind the observations is typically not known and hence selected from the sample. Note that this differs from HARKing. Let us consider an example. In a regression context, you may wish to test the null hypothesis that x affects y with a

coefficient equal to 1, $\beta = 1$ say. While this hypothesis is a priori, you may not know how many and which covariates z_1 through z_m to include to render a regression model ensuring valid inference about $\beta = 1$. Typically, the specification of the model is data driven. But even if the model selection step is consistent, it may affect and invalidate subsequent inference; see, for example, Leeb & Pötscher (2005); valid post-selection inference (PoSI) has been pioneered by Berk *et al*. (2013). To check whether the final model meets the assumptions required for valid inference, one may heed the advice by Hendry (1980, p. 430): 'The three golden rules of econometrics are test, test and test', although followed by a footnote saying 'Notwithstanding the difficulties involved in calculating and controlling type I and II errors'. In principle, carefully choosing and testing the model behind inference is well intentioned and not linked to *p*-hacking or MESSing. But standard model selection methods and multiple testing in this context more generally invalidate *p*-values. And hence, there may in practice be a smooth transition from model selection to MESSing: in the regression example, it is hard to rule out that the final specification was chosen such to produce a certain *p*-value when testing $\beta = 1$. Further, when multiple testing is not corrected for, it is a serious problem that invalidates inference (see, e. g. Farcomeni, 2008), but when some or all of the insignificant tests are not even reported, it turns into MESSing. Against this background, we consider it as problematic that strategies of model selection and specification search are covered in many textbooks without discussing that they invalidate subsequent inference.[4] We take Moore *et al*. (2014) as an influential example. Moore *et al*. (2014, sect. 11.2) include a multiple regression case study involving several steps: plot data, test hypotheses, check residuals and test more hypotheses. Indeed, Moore *et al*. (2014, p. 1079) highlight (with italics in the original): '*Multiple regression is a complicated procedure. If we do not do the necessary preliminary work, we are in serious danger of producing useless or misleading results*.' Of course, HARKing, *p*-hacking and other forms of data manipulation are not taught in textbooks, but as we have just argued, the transitions from such model selection strategies may be smooth. Thus, we believe that it is crucial that statistical education clearly draws a distinction between exploratory and confirmatory data analyses. Further, in our opinion, questionable research practices like HARKing or *p*-hacking and related issues like inference after model selection and multiple testing and the detrimental consequences that they all may have on the validity of statistical inference should be discussed in statistics courses from the introductory level on.

4 Small Manipulations, Big Effects

We propose a simple theoretical example that illustrates MESSing in both directions, that is, strengthening as well as weakening evidence subject to snooping and that makes clear that one is in a sense the mirror image of the other. We then use it to show how large the detrimental effects of these practices can be in terms of size and power distortions even in this simplistic setup.

Consider three researchers interested in assessing Gaussianity of daily stock returns. The first researcher hopes to find a significant deviation from Gaussianity to increase chances for publication of his article on non-normal return distributions. The second researcher is funded by the financial industry and may have no interest in higher capital requirements due to the increased risk implied by, for example, fat tails of the return distribution and thus is looking for insignificant results. The third researcher is just interested in scientific progress. All researchers use the same data set with n daily stock returns $x_1, x_2, \ldots, x_n$. We assume for simplicity that the returns are independent and consider three common and closely related tests of the null of Gaussianity based on the skewness of normal random variables being 0, $\gamma_1 = 0$, and the kurtosis being 3,

$\gamma_2 = 3$. The tests essentially assess the deviations of the empirical analogues of these standardised moments, $\hat{\gamma}_1$ and $\hat{\gamma}_2$, from 0 and 3, where

$$\hat{\gamma}_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^{k+2}}{d^{k+2}},\ k = 1,2 \text{ with } d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The first two tests use only one of the two moments, see also Shapiro *et al*. (1968), while the third omnibus test uses both, see Bowman & Shenton (1975, p. 243). The latter procedure is often called Jarque–Bera test after Jarque & Bera (1980). For the skewness and kurtosis tests, we use the standardised squares,

$$\Gamma_1^2 = n\frac{\hat{\gamma}_1^2}{6} \text{ and } \Gamma_2^2 = n\frac{(\hat{\gamma}_2 - 3)^2}{24},$$

respectively, which asymptotically follow a $\chi^2(1)$-distribution under the null, and the Jarque–Bera test uses the sum of the squares, $JB = \Gamma_1^2 + \Gamma_2^2$, which follows a $\chi^2(2)$-distribution asymptotically under Gaussianity.

The third researcher uses the full information, that is, the JB statistic, to test at level α and thus rejects the null if $JB > \chi^2_{1-\alpha}(2)$. The first researcher, striving to maximise evidence, checks first and secretly which sample moment violates the null most and then picks the skewness or kurtosis test accordingly, that is, determines $\Gamma^2_{\max} := \max_{k=1,2}(\Gamma_k^2)$ and rejects if $\Gamma^2_{\max} > \chi^2_{1-\alpha}(1)$. The second researcher, striving to minimise evidence, checks first which sample moment violates the null least and then picks the test in his favour, that is, determines $\Gamma^2_{\min} := \min_{k=1,2}(\Gamma_k^2)$ and rejects if $\Gamma^2_{\min} > \chi^2_{1-\alpha}(1)$.

The MESSing executed by the first and second researchers distorts the properties of the tests, which we quantify now. To calculate the sizes of $\Gamma^2_{\min}$ and $\Gamma^2_{\max}$ tests, we assume for simplicity that we have a rather large sample such that the underlying three tests have an actual size equal to the nominal size α (under normality):

$$\mathrm{P}(\Gamma_2^2 > \chi^2_{1-\alpha}(1)) = \mathrm{P}(\Gamma_1^2 > \chi^2_{1-\alpha}(1)) = \mathrm{P}(JB > \chi^2_{1-\alpha}(2)) = \alpha.$$

Let $R^{(k)}$ be the event that test k yields a significant result, that is, $\mathrm{P}(R^{(k)}) = \mathrm{P}(\Gamma_k^2 > \chi^2_{1-\alpha}(1)) = \alpha, k = 1,2$. It then holds (because these events are independent under Gaussianity) that

$$\begin{aligned}
\mathrm{P}(\Gamma^2_{\max} > \chi^2_{1-\alpha}(1)) &= \mathrm{P}(R^{(1)} \cup R^{(2)}) \\
&= \mathrm{P}(R^{(1)}) + \mathrm{P}(R^{(2)}) - \mathrm{P}(R^{(1)})\mathrm{P}(R^{(2)}) \\
&= 2\alpha - \alpha^2, \\
\mathrm{P}(\Gamma^2_{\min} > \chi^2_{1-\alpha}(1)) &= \mathrm{P}(R^{(1)} \cap R^{(2)}) \\
&= \mathrm{P}(R^{(1)})\mathrm{P}(R^{(2)}) = \alpha^2.
\end{aligned}$$

Table 1. *Simulated sizes and powers of the normality tests for a level of $\alpha = 0.05$, a sample size of 1000 and a t(10)-distribution under the alternative*

Test statistic	Γ_1^2	Γ_2^2	JB	$\Gamma^2_{\max}$	$\Gamma^2_{\min}$
Size	0.0495	0.0460	0.0485	0.0910	0.0046
Power	0.2782	0.9560	0.9405	0.9594	0.2747

For the nominal size of $\alpha = 5\%$, the actual sizes of $\Gamma^2_{\max}$ and $\Gamma^2_{\min}$ are 9.75% and 0.25%, respectively; see also Table 1 for finite sample evidence. Thus, MESSing through choosing a test in favour of the researchers' targets can almost double the size or let it almost disappear, and weakening evidence subject to snooping is as harmful as strengthening.

To illustrate the effect on power, we assume that the null is wrong in the direction of a fat-tailed alternative, namely, a *t*-distribution with 10 degrees of freedom, $t(10)$. We use a significance level of $\alpha = 0.05$ and a sample size of 1000 when simulating sizes and powers summarised in Table 1; for details, see the appendix. The $t(10)$-distribution is symmetric and leptokurtic ($\gamma_1 = 0$ and $\gamma_2 = 4$). As the deviation from normality is with respect to the fourth moment only, the kurtosis test has a very high power and the skewness test has a low power, $\mathrm{P}(R^{(2)}) = 0.9560$ and $\mathrm{P}(R^{(1)}) = 0.2782$, while the JB-test has a power of $\mathrm{P}(JB > \chi^2_{1-\alpha}(2)) = 0.9405$. Consequently, weakening evidence by essentially picking the skewness test is very effective here, $\mathrm{P}(\Gamma^2_{\min} > \chi^2_{1-\alpha}(1)) = 0.2747$. Strengthening evidence leads to a slight increase of the already high power of the JB-test used by the honest researcher, $\mathrm{P}(\Gamma^2_{\max} > \chi^2_{1-\alpha}(1)) = 0.9594$.

In this example, we allow for only one researcher degree of freedom, namely, the choice of the test statistic. Still in this simple case, Table 1 demonstrates that MESSing in both directions may already have drastic consequences, both under the null and alternative hypotheses. As researchers usually have many degrees of freedom (see Simmons *et al*., 2011), the example hints at how serious the consequences of MESSing may be.

5 Concluding Remarks

Data mining is a useful and essential tool to cope with the challenges of growing capacities to store and process massive amounts of data. This is not what this paper is about. We rather wish to stress a potential downside of data snooping in connection with statistical inference. Our empirical exercise with data from the German lottery and the theoretical example on testing for normality reinforce how misleading statistical hypothesis testing subsequent to data snooping can be and in how many different forms MESSes (manipulations of evidence subject to snooping) may come along. We also review the literature on questionable research practices, discussing several forms of MESSing and measures that have been put forward to discourage or prevent their use.

ACKNOWLEDGEMENTS

We are grateful to Matei Demetrescu, Steffen Eibelshäuser, Mehdi Hosseinkouchack, Michael Neugart, Jan Reitz and seminar participants at Free University Berlin for many helpful comments. Further, we thank two anonymous reviewers for helpful suggestions. Open access funding enabled and organized by Projekt DEAL.

Notes

[1]The data were downloaded from https://www.lotto.de/lotto-6aus49/statistik/ziehungshaeufigkeit on 29 November 2019.

[2]Note that standard results from order statistics do not apply here because S_1 through S_{49} are dependent. In general, it is a nontrivial problem to obtain distributional results for minima or maxima; see, for example, Nadarajah & Kotz (2008) for the case of two correlated Gaussian random variables. Thus, we use simulations here.

[3]On 17 June 1956, the drawing of the 7th additional number was introduced, and abolished on 4 May 2013.

[4]We are grateful to an anonymous referee for pointing this out.

[5]R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[6]https://stat.ethz.ch/R-manual/R-devel/library/stats/html/bandwidth.html.

References

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. 2013. Valid post-selection inference. *Ann. Stat.*, **41**(2), 802–837.

Bowman, K.O. & Shenton, L.R. 1975. Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, **62**, 243–250.

Brodeur, A., Cook, N. & Heyes, A. 2020. Methods matter: P-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.*, **110**(11), 3634–60.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. 2016. Star wars: The empirics strike back. *Am. Econ. J.: Appl. Econ.*, **8**(1), 1–32.

Christensen, G., Freese, J. & Miguel, E. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press.

Chuard, P.J.C., Vrtílek, M., Head, M.L. & Jennions, M.D. 2019. Evidence that nonsignificant results are sometimes preferred: Reverse p-hacking or selective reporting? *PLoS Biol.*, **17**(1), e3000127.

Coulson, M., Healey, M., Fidler, F. & Cumming, G. 2010. Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Front. Psychol.*, **1**, 26.

Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.*, **17**(4), 347–388.

Fricker Jr., R.D., Burke, K., Han, X. & Woodall, W.H. 2019. Assessing the statistical analyses used in Basic and Applied Social Psychology after their p-value ban. *Am. Statistician*, **73**(sup1), 374–384.

Genest, C., Lockhart, R.A. & Stephens, M.A. 2002. χ^2 and the lottery. *J. R. Stat. Soc.: Ser. D (The Statistician)*, **51**(2), 243–257.

Gerber, A. & Malhotra, N. 2008a. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quart. J. Polit. Sci.*, **3**(3), 313–326.

Gerber, A. & Malhotra, N. 2008b. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociolog. Methods Res.*, **37**(1), 3–30.

Hand, D.J. 1998. Data mining: Statistics and more? *Am. Stat.*, **52**(2), 112–118.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2. Springer.

Hendry, D.F. 1980. Econometrics–Alchemy or science? *Economica*, **47**(188), 387–406.

Hollenbeck, J.R. & Wright, P.M. 2017. Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *J. Manag.*, **43**(1), 5–18.

Imbens, G.W. 2021. Statistical significance, p-values, and the reporting of uncertainty. *J. Econ. Perspect.*, **35**(3), 157–74.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Med.*, **2**(8), e124.

Jarque, C.M. & Bera, A.K. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.*, **6**(3), 255–259.

Joe, H. 1993. Tests of uniformity for sets of lotto numbers. *Stat. Probab. Lett.*, **16**(3), 181–188.

John, L.K., Loewenstein, G. & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psycholog. Sci.*, **23**(5), 524–532.

Kerr, N.L. 1998. HARKing: Hypothesizing After the Results are Known. *Personality Social Psychol. Rev.*, **2**(3), 196–217.

Leeb, H. & Pötscher, B.M. 2005. Model selection and inference: Facts and fiction. *Econometr. Theory*, **21**(1), 21–59.

Lovell, M.C. 1983. Data mining. *Rev. Econ. Stat.*, **65**(1), 1–12.

Moore, D.S., McCabe, G.P. & Craig, B.A. 2014. *Introduction to the Practice of Statistics*, 8. Freeman.

Nadarajah, S. & Kotz, S. 2008. Exact distribution of the max/min of two Gaussian random variables. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, **16**(2), 210–212.

Nelson, L.D., Simmons, J. & Simonsohn, U. 2018. Psychology's renaissance. *Ann. Rev. Psychol.*, **69**, 511–534.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Mag. (Ser. 5)*, **50**(302), 157–175.

Rao, J.N.K. & Scott, A.J. 1981. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *J. Am. Stat. Assoc.*, **76**(374), 221–230.
Romer, D. 2020. In praise of confidence intervals. In *AEA Papers and Proceedings*, Vol. **110**, pp. 55–60.
Rubin, M. 2017. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.*, **21**(4), 308–320.
Shapiro, S.S., Wilk, M.B. & Chen, H.J. 1968. A comparative study of various tests for normality. *J. Am. Stat. Assoc.*, **63**(324), 1343–1372.
Simmons, J.P., Nelson, L.D. & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psycholog. Sci.*, **22**(11), 1359–1366.
Simonsohn, U. 2014. Posterior-hacking: Selective reporting invalidates Bayesian results also. Available at SSRN 2374040.
Simonsohn, U., Nelson, L.D. & Simmons, J.P. 2014. P-curve: A key to the file-drawer. *J. Experimental Psychol.: Gen.*, **143**(2), 534–547.
Sterling, T.D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *J. Am. Stat. Assoc.*, **54**(285), 30–34.
Sterling, T.D., Rosenbaum, W.L. & Weinkam, J.J. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *Am. Statistician*, **49**(1), 108–112.
Trafimow, D. & Marks, Michael 2015. Editorial. *Basic Appl. Social Psychol.*, **37**(1), 1–2.
Wallis, W.A. 1942. Compounding probabilities from independent signficance tests. *Econometrica*, **10**(3/4), 229–248.
Wasserstein, R.L. & Lazar, N.A. 2016. The ASA statement on p-values: Context, process, and purpose. *Am. Statistician*, **70**(2), 129–133.
Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. 2019. Moving to a world beyond "$p < 0.05$". *Am. Statistician*, **73** (sup1), 1–19.
White, H. 2000. A reality check for data snooping. *Econometrica*, **68**(5), 1097–1126.
White, J. & Bero, L.A. 2010. Corporate manipulation of research: Strategies are similar across five industries. *Stanford Law Policy Rev.*, **21**(1), 105–133.

Appendix A: Limiting Normality Under Dependence The general case of Lotto K out of M consists of K balls drawn without replacement in one game from an urn of M balls. Let $L_1, \ldots, L_K, L_{K+1}, \ldots, L_{2 \cdot K}, L_{2 \cdot K+1}, \ldots, L_{N \cdot K}$ be the consecutive numbers drawn in N games. To execute a test of uniformity from N games, we are interested in the counts of the M numbers from the sample of size $n = N \cdot K$. Let the counts of these numbers be denoted by S_m, $m = 1, 2, \ldots, M$, and consider the Bernoulli random variables

$$X_{m,i} = \begin{cases} 1 & \text{if } L_i = m \\ 0 & \text{if } L_i \neq m \end{cases}, \; i = 1 \ldots, N \cdot K = n,$$

which indicate if the i-th ball drawn shows the number m or not. These are the ingredients to determine the total counts $S_m = \sum_{i=1}^{n} X_{m,i}$. Under the null hypothesis of uniformity, $P(X_{m,i} = 1) = 1/M$ for all $m \in \{1, 2, \ldots, M\}$, and $X_{m,i} \sim Be(1/M)$ with $\mathrm{E}[S_m] = n/M$. However, due to the dependence between the Bernoulli random variables $X_{m,i}$ within one game caused by drawing without replacement, S_m does not follow a binomal distribution.

There is a simple way around the problem of the Bernoulli variates $X_{m,i}$ not being independent and consequently their sum S_m not being binomially distributed: one only requires a different standardisation to obtain a standard normal limit. Define for the j-th game the Bernoulli random variables indicating if the number m shows up in this game of K draws:

$$Y_{m,j} = \sum_{k=K(j-1)+1}^{Kj} X_{m,k}, \; j = 1, \ldots, N.$$

These variables are independent Bernoulli variates, $Y_{m,j} \sim Be(K/M)$, and by construction, they

determine the total counts: $S_m = \sum_{j=1}^{N} Y_{m,j}$. Therefore, S_m obeys the following binomial distribution: $S_m \sim Bi(N, K/M)$. The binomial test statistic hence becomes

$$Z_m^{lot} := \frac{S_m - \frac{K \cdot N}{M}}{\sigma_{lot}} \text{ with } \sigma_{lot}^2 := \frac{KN(M-K)}{M^2}, \tag{A1}$$

and by a classic central limit theorem, it holds under the null hypothesis that

$$Z_m^{lot} \xrightarrow{d} \mathcal{Z} \sim \mathcal{N}(0,1), \quad m = 1, \ldots, M. \tag{A2}$$

Appendix B: Monte Carlo Experiments Here, we provide details on the simulations used throughout the paper. All codes, which are written in R,[5] as well as the Lotto data are made available with this paper.

For the plot of the density in Figure 2 and to calculate the size distortions of the test, which uses standard normal critical values for the test statistic Z_m^{lot}, we approximated the distribution of $\mathcal{Z}_{\min}$ under uniformity of Lotto numbers by simulations. We simulated 10^4 times a sample of $6 \cdot 10^4$ Lotto numbers and determined the empirical distribution of the test statistics $\mathcal{Z}_{\min}$. The density, which is drawn in the picture, is a kernel density estimate, where the Gaussian kernel was used and Silverman's rule of thumb for bandwidth selection [see the documentation of the R Stats package (version 4.1.0)[6]].

For Table 1, we simulated 10^6 samples each of size $n = 10^3$ from a standard normal distribution under the null and from a t-distribution with 10 degrees of freedom under the alternative and executed the tests described in Section 4. The table contains the frequencies of rejection when testing at the 5% level.

[*Received March 2021; accepted January 2022*]