Adaptive Human Behavior and Physiology (2017) 3:43–60
DOI 10.1007/s40750-016-0050-z

METHODOLOGICAL CONTRIBUTIONS

To Remove or not to Remove: the Impact of Outlier Handling on Significance Testing in Testosterone Data

Thomas V. Pollet[1] **· Leander van der Meij**[1]

Received: 27 October 2015 / Revised: 28 June 2016 / Accepted: 5 July 2016 /
Published online: 29 August 2016

Abstract Outlier removal is common in hormonal research. Here we investigated to what extent removing outliers in hormonal data leads to divergent statistical conclusions. We first show that the most common outlier detection rule is based on a number of standard deviations (SD) from the mean. Next, we used simulations to examine the degree to which statistical conclusions diverge when a test with outlier exclusion yields a statistically significant result whereas the test with outlier inclusion did not, or vice versa (at $p = .05$). Simulations were run in duplicate for independent samples *t*-tests and repeated measures ANOVA designs, and based on real testosterone (T) data and a theoretical gamma distribution of T data. We ran simulations for different sample sizes (30 to 100) and outlier removal rules (2.5 SD and 3 SD). For significant *t*-tests, we found that in between 14 % to 55 % of the significant cases a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa (median *p* difference: .03–.06). For significant repeated measures ANOVAs, we found that in between 7 % to 28 % of significant cases a test where outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa (median *p* difference: .01–.03). When reporting any test that would lead to a statistically significant result (either the test with inclusion or exclusion of outliers (or both)), in between 5.15 % and 6.89 % of the independent sample *t*-tests were statistically significant, and for the repeated measures ANOVA design this was between 6.32 % and 7.62 % of the tests. Our results suggest that outlier handling can have a substantial impact on significance testing. We suggest several potential solutions for

Electronic supplementary material The online version of this article (doi:10.1007/s40750-016-0050-z) contains supplementary material, which is available to authorized users.

✉ Thomas V. Pollet
t.v.pollet@vu.nl

Leander van der Meij
l.van.der.meij@vu.nl

[1] Department of Experimental and Applied Psychology, VU University Amsterdam, Transitorium Building, room 1B17, 1081BT, Amsterdam, The Netherlands

handling outliers and we argue for a careful assessment of handling outliers in hormonal data.

Keywords Sex hormones · Statistical design · *p* value · Outlier handling · Statistical simulation

Introduction

Various scientific disciplines regularly come across 'outliers' in their data (e.g., Aguinis et al. 2013; Bakker and Wicherts 2014). Outliers are commonly defined as observations which are different from the majority of other cases in a sample (Barnett 1978; Barnett and Lewis 1994; Grubbs 1969; Hawkins 1980; Orr et al. 1991; Osborne and Overbay 2004; Rousseeuw and Hubert 2011). Such divergence could be due to, for example: measurement or coding error, sampling from the wrong population, exceptional circumstances, or a poor fit of the statistical model. Outliers matter and there are several famous cases where even a single case heavily impacts the results (e.g., Hollenbeck et al. 2006). For example, the journal *Acta Crystallographica A* had a stable impact factor of around 2.4 up until 2009 when it reached an impact factor of 49.9 due to a single paper with 5624 citations (Dimitrov et al. 2010). There is a very large body of literature on outlier handling, and an entire field in statistical pattern recognition is devoted to outlier detection and handling (e.g., Ritter and Gallegos 1997). Therefore, the purpose of this short report is not to review the importance and impact of outliers or all possible methods to deal with outliers (for in depth reviews see, for example: Aguinis et al. 2013; Barnett and Lewis 1994; Dixon 1953; Grubbs 1969; Hawkins 1980; Osborne and Overbay 2004; Rousseeuw and Hubert 2011; Shiffler 1988). Rather, we wish to examine the degree to which outlier removal decisions would impact the statistical conclusions in the context of endocrinological studies, and specifically those dealing with testosterone (T).

Outliers seem especially relevant for endocrinological studies, since hormonal data, such as T, typically do not conform to normal distributions. For example, due to its skewed distribution, T usually contains multiple outliers when defining outliers based on a normal distribution (Fig. 1 below; Stanton 2011: Fig. 1). Inclusion or exclusion of these outliers could matter, especially given the typically modest sample sizes in this research area. We therefore investigated to what extent removing outliers leads to different statistical conclusions based on statistical significance (p = .05). Our analyses focused on T, since this hormone is among the most frequently studied hormones and we had sufficient data available on T.

First, we reviewed outlier handling for all the articles in *Psychoneuroendocrinology, Hormones and Behavior,* and *Biological Psychology* from January 1st 2010 to May 1st 2015 that included some measure of T levels in humans (ESM Table 1-2-3). In 43 out of 133 papers (32.33 %) outlier handling was reported, and this mostly consisted of removal (35 out of 43). Occasionally winsorizing was used, which involves reassigning the values of extreme cases to, for example, the 95th percentile score (Ghosh and Vogt 2012; Hastings Jr et al. 1947). However, the most

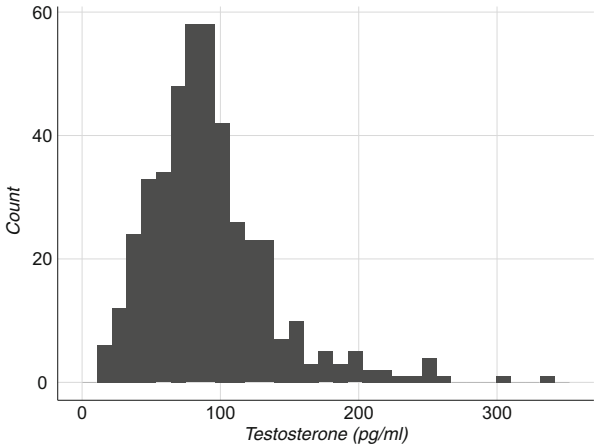

Fig. 1 Histogram of 433 testosterone samples

commonly applied rule for defining an outlier (24 out of 43 papers), was based on a case being a given number of standard deviations (SD) away from the raw mean scores.

It is thus clear that outliers are often reported and that a commonly applied procedure is to remove them. Here we investigated how often statistical conclusions based on 'p = .05 significance' diverged when a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa. We also assessed by how much p values differed when these statistical conclusions did differ. Finally, we also investigated how many tests were significant when reporting any result that was statistically significant (either the test with inclusion or exclusion of outliers). To this end, we simulated an independent samples t-test design and a pre-post repeated measures ANOVA design with varying sample sizes (30 to 100) and outlier removal rules (2.5 SD and 3 SD). Simulations were performed in duplicate on real hormonal data and on a theoretical gamma distribution of T data.

Methods

Data

T values were obtained from 433 saliva samples of men (7 unpublished studies by the authors, as well as one published study (van der Meij et al. 2015). All samples were collected in a similar fashion via the same protocol (for details: van der Meij et al. 2015) and were analyzed by the laboratory of Biological Psychology at the Dresden University of Technology. T was determined via an expanded range salivary T enzyme-immunoassay kit (cat. nu 1–2402) from Salimetrics (Suffolk, UK). The intra-and inter-assay coefficients are below 10 % and 12 %. No outliers were removed from the database (based on a 2.5 and 3 SD rule there would be 13 and 9 outliers, respectively).

Simulations

Simulations Based on Actual Data - Independent Samples T-Test Scenario

Using R 3.0.2 (R Development Core Team 2008) (packages used: Meyer et al. 2015; Wickham 2011, 2009, 2007, see ESM 4). We constructed a customized script to examine how often the inclusion or exclusion of an outlier would lead to different statistical conclusions at $p = .05$. Moreover, we investigated by how much these statistical conclusions would differ in terms of p value. The annotated script and data are included in ESM 4. The script was run in duplicate (Run 1 and 2) and worked as follows. First, it drew a sample of n (with replacement, $n = 30$–100) from our data ($N = 433$ T measurements; Fig. 1). Second, the script log transformed the data (log10) (e.g., Dabbs et al. 1995; Feldman et al. 2002; Pollet et al. 2013; van der Meij et al. 2008). Third, it then generated two arbitrary conditions of equal size by dividing the data into two groups. Fourth, the script performed two independent samples t-tests with the two previously constructed groups as independent variable: one t-test without removing any outliers and one t-test with *all* outliers removed. Outliers were defined using the raw data either as 2.5 SD or more or 3 SD or more away from the mean. Independent samples t-tests were corrected for unequal variances (Welch's t-test, see Ruxton 2006; Zimmerman 2004). All p values were two-tailed and we used a 5 % significance threshold ($p = .05$). Fifth, the script ran steps one to four 100,000 times. Each of those 100,000 cases can be interpreted as a different study (albeit constrained that they were sampled from our $n = 433$ with replacement). Sixth, it counted the number of significant t-tests with and without outlier removal. The script also calculated in how many of the cases the statistical conclusions based on significance differed: a test with outlier exclusion yielded a statistically significant result but the test with outlier inclusion did not, or vice versa (at $p = .05$). To examine this we thus have a 100,000 (no. of cases) × 2 matrix (outlier exclusion or inclusion). In order to calculate the % divergence the script summed up all the cases for which the statistical conclusions differed – i.e. one row entry in the matrix was significant but the other row entry was not (at $p = .05$ level). Next, this sum was divided by the total number of significant tests, which was then multiplied by a hundred to obtain a percentage (%). Thus, this measure could theoretically range between 0 % and 100 %. Zero percent divergence meant perfect correspondence in statistical conclusions, i.e., if a case was significant when excluding outliers, then this case was always also significant when including outliers, and vice versa. Hundred percent divergence meant zero correspondence in statistical conclusion, i.e., if a case was significant when excluding outliers this case was never significant when including outliers, and vice versa (at $p = .05$). Finally, the script calculated how much the p values differed in absolute terms for all cases for which the statistical conclusions based on significance differed. We reported the median of those absolute differences in p values as well as the 2.5 and 97.5 percentile of those differences in absolute p values. We also reported how many of those 100,000 entries were statistically significant at $p = .05$ if either the test with inclusion or exclusion of outliers was selected (note that this also counts the cases where both entries are significant).

Simulations Based on Actual Data – Repeated Measure ANOVA Scenario (RM-ANOVA)

As with the independent sample *t*-tests scenario using R 3.0.2 (R Development Core Team 2008) we constructed a customized script to examine examine how often the inclusion or exclusion of an outlier would lead to different statistical conclusions at $p = .05$ (ESM 4) for a RM-ANOVA design. Our model corresponds to a pre-post design with participants being assigned to one of two conditions. The script was run in duplicate (Run 1 and 2) and worked as follows. First, it drew a sample of n (30–80) with replacement (T1) from our data ($N = 433$ T measurements; Fig. 1). Second, the script used a Cholesky transformation matrix to generate correlated Time 2 (T2) data for each T1 data point (Stevenson 2013). To this end we first sampled T2 data from our sample ($N = 433$) with replacement, as we did for T1. This matrix with T1 and T2 matrix can then be transformed to obtain correlated values, with a certain correlation between T1 and T2. The matrix algebra is described at length in Golub and Van Loan 2012. In R, the command 'chol(matrix)' provides the transformation, and such a Cholesky transformation is an established way to create correlated data for simulations in R (for Matlab: (Van den Berg n.d..), for SAS: (SAS 2009; UCLA: Statistical Consulting Group 2016)). We chose to let this procedure create T2 values that were correlated with T1 with a ρ value of .5, since in a sample of our unpublished data, T1 and T2 were correlated with roughly this value ($r_{69} = .551$). However, the correlation can vary substantially across experimental designs, due to the time between measurements (samples closer together in time correlate more). These variations are partly accounted for in our analyses as the actual (sample) correlation for a single simulation can vary substantially from the population level correlation. Third, two arbitrary conditions of equal size were generated by arbitrarily assigning each 'participant' to one of these conditions (each participant had a T1 and T2 data point). Fourth, the script performed two Repeated Measures ANOVAs: one with inclusion of outliers and one with exclusion of outliers (based on a 2.5 and 3 SD outlier removal rule, see 2.2.1). For both scenarios, the script calculated the p value (two-tailed) of the interaction effect (between-within interaction: condition*time). Fifth, this process was reiterated 100,000 times. Sixth, it counted the number of significant *t*-tests with and without outlier removal. The script also calculated in how many of the cases the statistical conclusions based on significance differed: a test with outlier exclusion yielded a statistically significant result but the test with outlier inclusion did not, or vice versa (at $p = .05$). As with the independent samples *t*-test scenario, we thus examined a 100,000 (no. of cases) × 2 matrix (outlier exclusion or inclusion). In order to calculate the % divergence the script summed up all the cases for which the statistical conclusions differed – i.e. one row entry in the matrix was significant but the other row entry was not (at $p = .05$ level). Next, this sum was divided by the total number of significant tests, which was then multiplied by a hundred to obtain a percentage (%). Thus, this measure could theoretically range between 0 % and 100 %. Zero percent divergence meant perfect correspondence in statistical conclusions, i.e., if a case was significant when excluding outliers, then this case was always also significant when including outliers, and vice versa. Hundred percent divergence meant zero correspondence in statistical conclusion, i.e., if a case was significant when excluding outliers this case was never significant when including outliers, and vice versa (at $p = .05$). Finally, the script calculated how

much the p values differed in absolute terms for all cases for which the statistical conclusions based on significance differed. We reported the median of those absolute differences in p values as well as the 2.5 and 97.5 percentile of those differences in absolute p values. As with the independent samples t-test scenario, we also reported how many of those 100,000 entries were statistically significant at $p = .05$ if either the test with inclusion or exclusion of outliers was selected (note that this also counts the cases where both entries are significant).

Simulations Based on Theoretical Data

A drawback was that our resampling procedure was based on actual data and is thus also inherently limited to only those values. We therefore also sought to describe the distribution of T values and use simulations derived from a theoretical distribution rather than the actual values. Given the shape of the T distribution (Fig. 1), we examined several candidate distributions (normal, lognormal, Weibull and gamma distribution, (MASS package in R) (Ripley et al. 2015)). A gamma distribution was the best fit to the data based on the Kolmogorov-Smirnov tests ('ks.test' in R: ks.test(gamma) = 0.05, $p = .15$; ks.test(normal) = 1, $p < .0001$; ks.test(Weibull) = 0.086, $p = .003$; ks.test (lognormal) = 0.069, $p = .034$). Given that the ks.test(gamma) was not statistically significant, we maintained the null hypothesis that our observed distribution resembled the theoretical distribution. However, it is important to note that the Kolmogorov-Smirnov test for larger samples is a (very) conservative test, so other candidate distributions aside from gamma could have been unfairly rejected.

We thus ran the scripts explained in section 2.2.1 and 2.2.2 but instead of drawing from the 433 T values the script drew samples from the gamma distribution. Note that the theoretical distribution was generated based on our sample. Therefore, the 'true' distribution of male hormone samples could still look different from the one we modeled.

Results

The key simulation results are summarized in Tables 1-4. To illustrate how the tables should be interpreted we interpret the first row of Run 1 in Table 1. With a sample size of $n = 30$, actual data, and a 2.5 SD outlier removal rule, the first simulation (Run 1) found a base rate significance of 4.85 % when outliers were included (No outlier α) and 4.72 % tests when all outliers were excluded (Outlier α). In 49.37 % of the significant cases, including outliers led to a statistically significant result whereas outlier exclusion led to a statistically non-significant result, or vice versa (at $p = .05$, see title column and also see description in method section). When the statistical conclusions differed, they differed by around 6 % in p value (median estimate, 95 % range: 1 % to 26 %), see title column. Figure 2 plots these divergent estimates (with the upper 2.5 % tail not shown, as the maximum p value is .84 (!)).

Across all simulations, the worst case scenario suggests that in over 50 % of the significant cases including outliers led to a statistically significant result whereas outlier exclusion led to a non- statistically significant result, or vice versa, and if these results did differ, the median difference in p values was .07 (e.g., $p = .04$ vs. $p = .11$)

Table 1 The % of significant independent samples *t*-tests (at $p = .05$) with and without outlier removal for a given sample size (No outlier α, Outlier α). % divergent refers to the percentage of tests where the statistical conclusions differed, i.e. one test significant but another one not (at $p = .05$), as compared to all statistically significant entries. Median refers to the median (absolute) difference in *p* value where one test was significant but another one was not, and 2.5 % - 97.5 % give the corresponding percentiles. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on $N = 433$ testosterone samples

Actual Data - Independent Samples t-test

		Run 1				Run 2			
	N	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)
2.5 SD rule	30	4.85 %	4.72 %	49.37 %	0.06 (0.01–0.26)	4.69 %	4.68 %	49.72 %	0.06 (0.01–0.28)
	40	4.86 %	4.89 %	52.69 %	0.05 (0.01–0.28)	4.87 %	4.84 %	52.77 %	0.05 (0.01–0.29)
	50	4.93 %	4.87 %	54.44 %	0.05 (0.01–0.30)	5.01 %	4.93 %	55.02 %	0.05 (0.01–0.29)
	60	4.88 %	4.85 %	54.42 %	0.05 (0.01–0.31)	4.99 %	4.98 %	53.95 %	0.06 (0.01–0.32)
	70	4.91 %	4.92 %	53.09 %	0.06 (0.01–0.34)	4.95 %	4.90 %	54.20 %	0.06 (0.01–0.33)
	80	4.97 %	5.04 %	53.80 %	0.06 (0.01–0.33)	5.03 %	5.03 %	53.95 %	0.06 (0.01–0.35)
	90	4.97 %	4.90 %	54.40 %	0.06 (0.01–0.37)	5.02 %	4.96 %	53.69 %	0.06 (0.01–0.36)
	100	4.95 %	4.95 %	54.59 %	0.06 (0.005–0.36)	4.90 %	4.89 %	54.79 %	0.07 (0.005–0.37)
3 SD rule	30	4.85 %	4.80 %	32.00 %	0.06 (0.01–0.23)	4.69 %	4.70 %	31.89 %	0.07 (0.01–0.23)
	40	4.89 %	4.94 %	37.83 %	0.05 (0.01–0.21)	4.87 %	4.93 %	37.72 %	0.05 (0.01–0.21)
	50	4.93 %	4.88 %	42.52 %	0.05 (0.01–0.21)	5.01 %	4.94 %	42.50 %	0.05 (0.01–0.21)
	60	4.88 %	4.86 %	44.56 %	0.05 (0.01–0.23)	4.99 %	4.96 %	44.35 %	0.05 (0.01–0.22)
	70	4.91 %	4.88 %	45.86 %	0.04 (0.01–0.24)	4.95 %	4.95 %	45.95 %	0.05 (0.01–0.24)
	80	4.97 %	4.99 %	46.67 %	0.05 (0.01–0.24)	5.03 %	5.04 %	46.87 %	0.05 (0.01–0.25)
	90	4.97 %	4.93 %	47.25 %	0.05 (0.01–0.25)	5.02 %	5.00 %	47.07 %	0.05 (0.01–0.26)
	100	4.95 %	4.93 %	47.37 %	0.05 (0.01–0.25)	4.90 %	4.89 %	47.71 %	0.05 (0.01–0.26)

(depending on statistical test, *n,* outlier removal rule, sample size, and Run no.), see Tables 1-4 for the results. However, do note that for an individual case it could be substantially more than .07. If we take the upper estimate of the range, it could be as much as .37 difference in *p* value. That would imply a shift from, for example, $p = .04$ to $p = .41$. However, the best case scenario suggests, that the conclusions might only vary in 7 % of the cases, and if these cases do vary, they do so by .01 in *p* value.

Independent Samples *t*-Test

For the independent samples *t*-test scenario based on actual data, in between 32 % and 55 % of the significant cases a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa (at $p = .05$, depending on *n*, outlier removal rule, and Run no.). When these statistical conclusions did differ, they differed between .05 to .07 in terms of *p* value (overall range: .01 to .37). For the simulations based on the theoretical distribution, results

Table 2 The % of significant independent samples *t*-tests (at $p = .05$) with and without outlier removal for a given sample size (No outlier α, Outlier α). % divergent refers to the percentage of tests where the statistical conclusions differed, i.e. one test significant but another one not (at $p = .05$), as compared to all statistically significant entries. Median refers to the median difference in *p* value where one test was significant but another one was not, and 2.5 % - 97.5 % give the corresponding percentiles. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on a theoretical gamma distribution

Theoretical data - Independent Samples t-test

		Run 1				Run 2			
	N	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)
2.5 SD rule	30	4.81 %	4.79 %	33.46 %	0.05 (0.01–0.19)	4.79 %	4.79 %	34.14 %	0.05 (0.01–0.18)
	40	4.81 %	4.78 %	37.17 %	0.05 (0.01–0.19)	4.87 %	4.81 %	38.34 %	0.05 (0.01–0.19)
	50	4.86 %	4.83 %	41.73 %	0.04 (0.01–0.21)	4.85 %	4.92 %	42.28 %	0.04 (0.01–0.20)
	60	4.84 %	4.83 %	44.18 %	0.04 (0.01–0.20)	4.85 %	4.87 %	42.72 %	0.04 (0.01–0.20)
	70	5.00 %	4.92 %	42.51 %	0.04 (0.01–0.21)	4.98 %	4.94 %	43.79 %	0.04 (0.01–0.22)
	80	5.03 %	4.94 %	44.23 %	0.04 (0.01–0.23)	5.03 %	5.06 %	43.44 %	0.04 (0.01–0.21)
	90	5.05 %	5.00 %	43.86 %	0.05 (0.01–0.23)	5.06 %	5.13 %	43.12 %	0.04 (0.01–0.22)
	100	5.04 %	4.89 %	44.48 %	0.04 (0.01–0.23)	5.01 %	5.10 %	43.92 %	0.04 (0.01–0.23)
3 SD rule	30	4.81 %	4.78 %	14.74 %	0.06 (0.01–0.21)	4.79 %	4.76 %	14.60 %	0.06 (0.01–0.20)
	40	4.81 %	4.78 %	18.90 %	0.05 (0.01–0.16)	4.87 %	4.86 %	19.44 %	0.05 (0.01–0.16)
	50	4.86 %	4.82 %	23.49 %	0.04 (0.01–0.15)	4.85 %	4.86 %	24.49 %	0.04 (0.01–0.15)
	60	4.84 %	4.82 %	25.96 %	0.04 (0.01–0.15)	4.85 %	4.87 %	27.29 %	0.04 (0.01–0.13)
	70	5.00 %	4.95 %	27.17 %	0.05 (0.01–0.14)	4.98 %	4.97 %	28.12 %	0.04 (0.01–0.14)
	80	5.03 %	4.97 %	29.49 %	0.04 (0.01–0.15)	5.03 %	5.04 %	29.47 %	0.04 (0.01–0.14)
	90	5.05 %	5.05 %	29.51 %	0.03 (0.01–0.16)	5.06 %	5.07 %	30.20 %	0.03 (0.01–0.15)
	100	5.04 %	4.97 %	32.27 %	0.03 (0.01–0.16)	5.01 %	5.10 %	31.50 %	0.03 (0.01–0.14)

showed a smaller difference compared to the actual data: statistical conclusions based on inclusion versus exclusion of outliers differed between 15 % and 45 % of the significant cases, and when they did differ, they differed between .05 to .07 in terms of *p* value (overall range: .01 to .23, depending on *n*, outlier removal rule, and Run no.).

RM-ANOVA

Dependent on sample size, outlier removal rule, and Run no., for simulations based on actual data, in between 15 and 28 % of significant cases a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa ($p = .05$ level). For simulations based on theoretical distributions the statistical conclusions diverged between 7 and 20 % of significant cases. For those cases where the statistical conclusions diverged, the median divergence in *p* values was around .02–.03 for simulations based on actual data, and around .01–.02 for simulations based on theoretical distributions. The corresponding ranges (2.5 percentile and 97.5 percentile) were between <.01 to .09 for simulations based on actual data and between <.01 to .07 for simulations based on theoretical distributions.

Table 3 The % of significant tests (at $p = .05$) for the interaction in a Repeated Measures ANOVA (within-between) with and without outlier removal for a given sample size (No outlier α, Outlier α). % divergent refers to the percentage of tests where the statistical conclusions differed, i.e. one test significant but another one not (at $p = .05$), as compared to all statistically significant entries. Median refers to the median (absolute) difference in p value where one test was significant but another one was not, and 2.5 % - 97.5 % give the corresponding percentiles.. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on $N = 433$ testosterone samples

Actual data - repeated measures anova

		Run 1				Run 2			
	N	NO OUTLIER α	OUTLIER α	% DIVERGENT	MEDIAN (2.5 % - 97.5 %)	NO OUTLIER α	OUTLIER α	% DIVERGENT	MEDIAN (2.5 % - 97.5 %)
2.5 SD rule	30	6.63 %	6.63 %	25.92 %	0.03 (0.006–0.09)	6.71 %	6.59 %	24.61 %	0.03 (0.005–0.09)
	40	6.46 %	6.39 %	27.24 %	0.03 (0.006–0.08)	6.37 %	6.31 %	26.80 %	0.03 (0.005–0.09)
	50	6.22 %	6.20 %	26.99 %	0.03 (0.005–0.08)	6.35 %	6.41 %	27.30 %	0.03 (0.005–0.08)
	60	6.19 %	6.23 %	27.40 %	0.03 (0.005–0.08)	6.14 %	6.14 %	27.24 %	0.03 (0.006–0.08)
	70	6.22 %	6.18 %	27.61 %	0.03 (0.005–0.08)	6.01 %	6.08 %	27.60 %	0.03 (0.005–0.08)
	80	6.18 %	6.22 %	27.95 %	0.03 (0.005–0.08)	5.91 %	5.88 %	28.39 %	0.03 (0.005–0.08)
	90	6.02 %	6.02 %	27.65 %	0.03 (0.005–0.08)	6.02 %	6.08 %	28.20 %	0.03 (0.005–0.08)
	100	6.03 %	6.07 %	27.54 %	0.03 (0.006–0.08)	5.90 %	5.91 %	28.23 %	0.03 (0.005–0.08)
3 SD rule	30	6.63 %	6.60 %	16.04 %	0.02 (0.005–0.08)	6.71 %	6.67 %	15.29 %	0.02 (0.005–0.07)
	40	6.46 %	6.39 %	18.51 %	0.02 (0.005–0.07)	6.37 %	6.33 %	17.83 %	0.02 (0.005–0.07)
	50	6.22 %	6.20 %	18.88 %	0.02 (0.005–0.06)	6.35 %	6.34 %	19.30 %	0.02 (0.005–0.06)
	60	6.19 %	6.17 %	19.74 %	0.02 (0.004–0.06)	6.14 %	6.08 %	19.58 %	0.02 (0.004–0.06)
	70	6.22 %	6.14 %	20.28 %	0.02 (0.004–0.05)	6.01 %	6.06 %	20.10 %	0.02 (0.004–0.06)
	80	6.18 %	6.23 %	21.10 %	0.02 (0.004–0.05)	5.91 %	5.92 %	20.90 %	0.02 (0.004–0.06)
	90	6.02 %	6.02 %	20.79 %	0.02 (0.004–0.06)	6.02 %	6.03 %	20.27 %	0.02 (0.004–0.05)
	100	6.03 %	6.12 %	19.75 %	0.02 (0.004–0.05)	5.90 %	5.93 %	21.45 %	0.02 (0.004–0.05)

Table 4 The number of significant tests (at $p = .05$) for the interaction in a Repeated Measures ANOVA (within-between) with and without outlier removal for a given sample size (No outlier α, Outlier α). % divergent refers to the percentage of tests where the statistical conclusions differed, i.e. one test significant but another one not (at $p = .05$), as compared to all statistically significant entries. Median refers to the median difference in p value where one test was significant but another one was not, and 2.5 % - 97.5 % give the corresponding percentiles. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on a theoretical gamma distribution

Theoretical data - repeated measures anova

		Run 1				Run 2			
	N	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)	No outlier α	Outlier α	% Divergent	Median (2.5 % - 97.5 %)
2.5 SD rule	30	6.66 %	6.71 %	15.44 %	0.02 (0.004–0.07)	6.85 %	6.94 %	15.42 %	0.02 (0.004–0.07)
	40	6.30 %	6.41 %	17.13 %	0.02 (0.004–0.06)	6.34 %	6.43 %	16.96 %	0.02 (0.004–0.06)
	50	6.22 %	6.29 %	17.94 %	0.02 (0.004–0.06)	6.18 %	6.31 %	18.40 %	0.02 (0.004–0.06)
	60	6.14 %	6.33 %	18.51 %	0.02 (0.003–0.06)	6.01 %	6.12 %	19.27 %	0.02 (0.004–0.05)
	70	6.09 %	6.23 %	18.96 %	0.02 (0.003–0.05)	5.93 %	6.09 %	19.97 %	0.02 (0.003–0.05)
	80	5.95 %	6.11 %	19.13 %	0.02 (0.003–0.06)	6.05 %	6.14 %	18.46 %	0.02 (0.003–0.05)
	90	6.08 %	6.17 %	19.72 %	0.02 (0.004–0.05)	5.96 %	6.14 %	19.80 %	0.02 (0.004–0.05)
	100	5.92 %	6.02 %	19.90 %	0.02 (0.004–0.05)	6.07 %	6.16 %	18.30 %	0.02 (0.003–0.05)
3 SD rule	30	6.66 %	6.67 %	7.09 %	0.02 (0.004–0.06)	6.85 %	6.87 %	7.37 %	0.02 (0.004–0.05)
	40	6.30 %	6.35 %	8.64 %	0.02 (0.003–0.05)	6.34 %	6.36 %	8.45 %	0.02 (0.004–0.05)
	50	6.22 %	6.28 %	9.54 %	0.02 (0.004–0.04)	6.18 %	6.25 %	9.91 %	0.02 (0.003–0.05)
	60	6.14 %	6.28 %	10.36 %	0.02 (0.003–0.05)	6.01 %	6.09 %	10.48 %	0.02 (0.003–0.04)
	70	6.09 %	6.14 %	10.30 %	0.01 (0.003–0.04)	5.93 %	6.01 %	11.57 %	0.01 (0.003–0.04)
	80	5.95 %	6.06 %	11.94 %	0.01 (0.003–0.04)	6.05 %	6.07 %	10.94 %	0.01 (0.003–0.03)
	90	6.08 %	6.09 %	12.46 %	0.01 (0.002–0.04)	5.96 %	6.07 %	11.42 %	0.01 (0.003–0.04)
	100	5.92 %	5.95 %	12.20 %	0.01 (0.003–0.04)	6.07 %	6.11 %	12.22 %	0.01 (0.002–0.04)

Robustness and Sample Size

Results were generally quite stable between runs. The lowest correlation between runs was the RM-ANOVA scenario based on theoretical data (for proportions of divergent tests: $r = .85$ for 2.5 SD outlier removal rule, $r = .92$ for 3 SD outlier removal rule). For the remainder of the scenarios correlations between Run 1 and Run 2 for the proportions of divergent tests was >.94).

Results also showed that the proportion of divergent tests was larger in larger sample sizes than in smaller sample size. For the independent samples t-tests scenario the % of divergent tests increased with sample size. For example, with a 3 SD outlier removal rule, 32 % of the tests diverge with a sample size of 30, while it was around 47 % for a sample size of a 100 (see Table 1). The median divergence in p remained relatively stable at around 5–6 %. For the RM-ANOVA scenario simulations based on actual data, increasing sample size also led to a higher proportion of tests with divergent conclusions (Table 3: 3 SD outlier removal rule: from 15 to 21 %). A potential explanation for these findings is that in larger samples sizes outliers are more likely

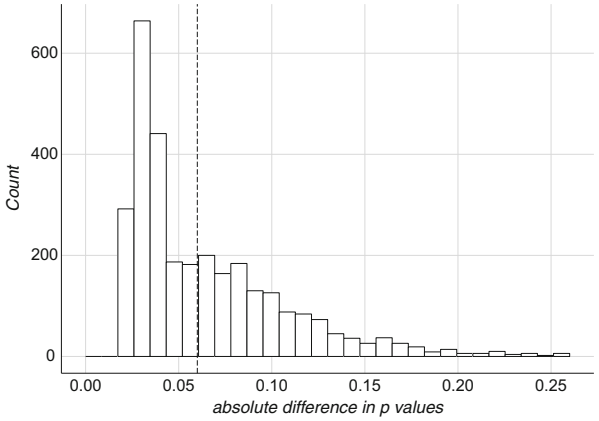

Fig. 2 Illustration of divergence: Histogram of absolute difference in p values when including or excluding outliers (Run 1, 2.5 SD rule, $n = 30$, actual data) and the statistical conclusions differ at $p = .05$. Dashed line is median estimate. Note: Upper 2.5 % of data not shown (as maximum is $p = .84$). Note that this is merely an illustration of just 1 entry in Table 1. The code is included in ESM 4 to draw histograms for other entries in Tables 1-4

to occur. In these larger samples there are thus more outliers to potentially remove and this makes that inclusion or exclusion of outliers could matter more often.

Furthermore, the statistical conclusions based on significance differed more often for the 2.5 SD outlier removal rule than for the 3 SD outlier removal rule (Table 1-4). For example, in the 2.5 SD outlier removal rule and t-test based on actual data, the conclusions differed in 44 % to 54 % of the cases (depending on n and Run no.), while for the 3 SD outlier removal rule the statistical conclusions differed in 32 % to 48 % of the cases. A potential explanation for these findings is that more outliers are defined using the 2.5 SD rule than the 3 SD rule, and this makes that inclusion or exclusion of outliers could matter more often. The corresponding median estimates for divergence in p value do not vary that substantially between the 2.5 SD and 3 SD rule. For example, for both the 2.5 SD and 3 SD rule, the median estimates for divergence were between .05 and .07 in p value for t-tests based on actual data.

Base Rate Significance

We also obtained baseline significance rates for each 100,000 tests. For the independent samples t-tests we found significance rates close to 5 % across all sample sizes (Tables 1-2: 4680 to 5104 out of 100,000 tests were statistically significant). However, the baseline significance rate was considerably higher for the RM-ANOVA scenario (Tables 3-4: 5884 to 6924 out of 100,000 tests were statistically significant). The 5.9 % to 6.9 % base rate suggests that Type I errors were inflated. A potential cause of this inflation is that the normality assumption of the RM-ANOVA was violated. This is suggested by the finding that some skew remained in our data even after log10 transforming the raw T scores (Skewness statistic: −.505; one-sample Kolmogorov-Smirnov test for log10(T) values: .07, $p = .03$).

When reporting any test that would lead to a statistically significant result (either the test with inclusion or exclusion of outliers), in between 5.15 % and 6.89 % of the

independent sample *t*-tests were statistically significant (Table 5-6; see description in method section), and for the repeated measures ANOVA design this was between 6.32 % and 7.62 % of the tests (Tables 7-8).

Discussion and Conclusion

Our results show that statistical conclusions based on significance can differ often and substantially if outliers are included or excluded. Our simulations showed that in between in 7 % to 54 % of significant cases a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa, depending on the statistical test, sample size, and outlier criterion. Our simulations should thus be taken as an illustration of how conclusions can diverge in endocrinological research.

Moreover, our results illustrate that inclusion or exclusion of outliers offer one option to '*p* hack' (Simmons et al. 2011). Researchers are thus able to shift statistical conclusions from non-significant to significant (or vice versa) by including or excluding outliers in a relatively high percentage of cases. Our results also showed that between 5.15 % and 7.62 % of the tests were statistically significant when reporting any test that yielded a statistically significant result (either inclusion or exclusion of

Table 5 Base rate significance of independent samples *t*-test when counting either test significant, i.e. one with or without outlier removal. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on $N = 433$ testosterone samples

		Actual Data - Independent Samples t-test	
	N	Run 1 α (when choosing either test)	Run 2 α (when choosing either test)
2.5 SD rule	30	6.35 %	6.23 %
	40	6.67 %	6.60 %
	50	6.73 %	6.86 %
	60	6.69 %	6.83 %
	70	6.69 %	6.76 %
	80	6.85 %	6.89 %
	90	6.77 %	6.82 %
	100	6.81 %	6.75 %
3 SD rule	30	5.74 %	5.59 %
	40	6.07 %	6.04 %
	50	6.23 %	6.32 %
	60	6.27 %	6.39 %
	70	6.35 %	6.43 %
	80	6.50 %	6.57 %
	90	6.48 %	6.55 %
	100	6.47 %	6.43 %

Table 6 Base rate significance of independent samples *t*-test when counting either test significant, i.e. one with or without outlier removal. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on a theoretical gamma distribution

		Theoretical Data - Independent Samples t-test	
		Run 1	Run 2
	N	α (when choosing either test)	α (when choosing either test)
2.5 SD rule	30	5.77 %	5.78 %
	40	5.89 %	5.99 %
	50	6.12 %	6.19 %
	60	6.20 %	6.18 %
	70	6.30 %	6.35 %
	80	6.40 %	6.45 %
	90	6.44 %	6.49 %
	100	6.39 %	6.48 %
3 SD rule	30	5.17 %	5.15 %
	40	5.30 %	5.39 %
	50	5.48 %	5.53 %
	60	5.55 %	5.63 %
	70	5.76 %	5.79 %
	80	5.87 %	5.91 %
	90	5.92 %	5.96 %
	100	5.97 %	6.00 %

outliers). In absolute terms this alpha inflation could be seen as small. However, if the statistical conclusions based on significance were at odds with one another, the difference in *p* value between the two conclusions was substantial in some cases. Consequently, outlier handling could be an (additional) important reason why some findings are difficult to replicate in behavioral endocrinology (e.g., Oxytocin: Nave et al. 2015). Thus, researchers, reviewers, and editors need to address outlier handling in hormonal data. However, if one is determined to obtain significant results, then other routes also exist such as, for example, running multiple tests (without correction) and only focusing on significant tests (Gelman and Loken 2013). Future research is necessary to determine the relative opportunities to *p* hack for various questionable research practices. Modeling several such questionable research practices, such as flexibility in outlier removal, would be an interesting future avenue for research. This could be done for example via *p* curves (Simonsohn et al. 2014).

How could we deal with these issues concerning outlier removal? We believe that a multitude of solutions exist. Firstly, we strongly recommend reporting the results both with outlier exclusion and inclusion, and if the conclusions differ strongly then caution is warranted (Kruskal 1960). Ideally this would be done graphically, so that one can really see the consequence of a given outlier. Secondly, researchers should also consider using non-parametric criteria for outlier detection. For example, using absolute deviations from the median rather than the mean (Leys et al. 2013), or relying on

Table 7 Base rate significance of Repeated-Measures ANOVA (within-between interaction) when counting either test significant, i.e. one with or without outlier removal. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on $N = 433$ testosterone samples

		Actual Data - Repeated Measures ANOVA	
		Run 1	Run 2
	N	α (when choosing either test)	α (when choosing either test)
2.5 SD rule	30	7.62 %	7.58 %
	40	7.43 %	7.32 %
	50	7.18 %	7.38 %
	60	7.20 %	7.11 %
	70	7.19 %	7.01 %
	80	7.21 %	6.87 %
	90	6.99 %	7.04 %
	100	7.01 %	6.88 %
3 SD rule	30	7.19 %	7.24 %
	40	7.08 %	6.97 %
	50	6.86 %	7.02 %
	60	6.86 %	6.78 %
	70	6.87 %	6.71 %
	80	6.94 %	6.60 %
	90	6.72 %	6.70 %
	100	6.74 %	6.63 %

interquartile range for outlier removal (Tukey 1977). A third option would be to not remove any 'outliers' at all (unless they are caused by a measurement error) and analyze the data with bootstrapping (Davison and Hinkley 1997), since in bootstrapping outliers receive proportionally less weight. Alternatively, robust statistics can also be used (Huber 2011; Rousseeuw and Hubert 2011; Rousseeuw and Leroy 2005). Fourth, setting a pre-defined criterion on outlier removal is also helpful. All these previous recommendations will have a greater impact if researchers make the raw data available, so that all outlier handling will be open, transparent and reproducible (Nosek et al. 2015). Finally, in the longer run, it would be a strong move forward, if we can combine baseline salivary data from various sources in order to generate population based reference values, as has been done for height (e.g., Wikland et al. 2002). Such a reference database can then meaningfully serve in decisions on whether or not to consider certain values as extreme or not.

In our review, we showed that the most common outlier detection rule was a rule based on a 2.5 or 3 SD difference from the mean. However, this is possibly not the best way to identify outliers in T data, as these data have skewed distributions. Consequently, more outliers are detected using an SD outlier removal rule in skewed data as opposed to in normally distributed data. As an example, in our dataset ($n = 433$), a 3 SD rule would allow removing up to 2.1 % of the data based on such an outlier criterion (9

Table 8 Base rate significance of Repeated-Measures ANOVA (within-between interaction) when counting either test significant, i.e. one with or without outlier removal. Results are presented for two runs and for outlier removal based on a 2.5 SD and 3 SD rule. These simulations were based on on a theoretical gamma distribution

		Theoretical Data - Repeated Measures ANOVA	
		Run 1	Run 2
	N	α (when choosing either test)	α (when choosing either test)
2.5 SD rule	30	7.25 %	7.47 %
	40	6.95 %	6.98 %
	50	6.87 %	6.88 %
	60	6.87 %	6.71 %
	70	6.80 %	6.67 %
	80	6.67 %	6.71 %
	90	6.79 %	6.71 %
	100	6.63 %	6.73 %
3 SD rule	30	6.91 %	7.12 %
	40	6.61 %	6.63 %
	50	6.56 %	6.54 %
	60	6.55 %	6.38 %
	70	6.45 %	6.34 %
	80	6.38 %	6.41 %
	90	6.49 %	6.38 %
	100	6.32 %	6.48 %

cases, all at the upper tail). This is a *fifteen-fold* increase of outliers: for a standard normal distribution this should be only 0.1 % (0.6 case, at upper tail). It is important to reiterate that in certain cases outlier removal may be fully warranted, for example, due to a measurement error. However, blindly applying an outlier removal rule that is based on a normal distribution would label some cases as outliers while they are not necessarily deviating from the 'true' T distribution.

There are some limitations to our study. For example, we limited ourselves to a single hormone, testosterone. Similarly, our starting point was 433 T samples, and therefore our simulations will be inherently limited by that sample. We also focused on a $p = .05$ criterion, rather than modeling the reduction of effect sizes. The reason for doing so, is that it is clear that p values, and a 5 % cutoff, still feature prominently in the literature (e.g., Bakker et al. 2012; Button et al. 2013; Ioannidis et al. 2014) and that many researchers, reviewers, and editors decide that findings are important if $p \leq .05$ (e.g., in psychophysics: Hoekstra et al. 2006; also see: Head et al. 2015; Leggett et al. 2013).

In conclusion, we believe that outlier handling is an issue that needs to be addressed in hormonal research since our simulations illustrate how statistical conclusions can vary substantially based on inclusion or exclusion of outliers. A first, achievable, step would be to encourage further openness in how results would differ if outliers are

excluded or included. A second step would be to create guidelines for both reporting and removing outliers in hormonal data (e.g., Aguinis and Edwards 2014; Aguinis et al. 2013). We hope our paper can serve as a first stepping-stone to a debate on outliers and their impact on statistical conclusions in the field of hormones and behavior.

Acknowledgments This work was supported in part by a NWO (Veni, 451.10.32) and a grant by the Netherlands Institute of Advanced Study in the humanities and Social Sciences (NIAS) to the first author and an internal faculty (equipment fund) grant to both authors.

References

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies, 51*, 143–174. doi:10.1111/joms.12058.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*, 270–301. doi:10.1177/1094428112470848.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychological Methods, 19*, 409–427. doi:10.1037/met0000014.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554. doi:10.1177/1745691612459060.

Barnett, V. (1978). The study of outliers: purpose and model. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 27*, 242–250. doi:10.2307/2347159.

Barnett, V., & Lewis, T., (1994). Outliers in statistical data. Wiley, New York, NY.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. doi:10.1038/nrn3475.

Dabbs, J. M., Campbell, B. C., Gladue, B. A., Midgley, A. R., Navarro, M. A., Read, G. F., Susman, E. J., Swinkels, L. M., & Worthman, C. M. (1995). Reliability of salivary testosterone measurements: a multicenter evaluation. *Clinical Chemistry, 41*, 1581–1584.

Davison, A.C., & Hinkley, D. V, (1997). Bootstrap methods and their application. Cambridge University Press, Cambridge, UK.

Dimitrov, J. D., Kaveri, S. V., & Bayry, J. (2010). Metrics: journal's impact factor skewed by a single paper. *Nature, 466*, 179. doi:10.1038/466179b.

Dixon, W. J. (1953). Processing data for outliers. *Biometrics, 9*, 74–89.

Feldman, H. A., Longcope, C., Derby, C. A., Johannes, C. B., Araujo, A. B., Coviello, A. D., Bremner, W. J., & McKinlay, J. B. (2002). Age trends in the level of serum testosterone and other hormones in middle-aged men: longitudinal results from the Massachusetts male aging study. *The Journal of Clinical Endocrinology and Metabolism, 87*, 589–598. doi:10.1210/jc.87.2.589.

Gelman, A., & Loken, E., (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition"or "p-hacking"and the research hypothesis was posited ahead of time [WWW Document]. URL http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Ghosh, D., & Vogt, A., (2012). Outliers: an evaluation of methodologies [WWW document]. Jt. Stat Meet URL https://www.amstat.org/sections/srms/proceedings/y2012/files/304068_72402.pdf

Golub, G.H., & Van Loan, C.F., (2012). Matrix computations. Johns Hopkins University Press, Baltimore, MD.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*, 1–21.

Hastings Jr, C., Mosteller, F., Tukey, J. W., & Winsor, C. P. (1947). Low moments for small samples: a comparative study of order statistics. *Annals of Mathematical Statistics, 18*, 413–426.

Hawkins, D.M., (1980). Identification of outliers. Springer, New York, NY.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*, e1002106. doi:10.1371/journal.pbio.1002106.

Hoekstra, R., Finch, S., Kiers, H. L., & Johnson, A. (2006). Probability as certainty: dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review, 13*, 1033–1037. doi:10.3758/BF03213921.

Hollenbeck, J. R., DeRue, D. S., & Mannor, M. (2006). Statistical power and parameter stability when subjects are few and tests are many: Comment on Peterson, Smith, Martorana, and Owens (2003). *The Journal of Applied Psychology, 91*, 1–5. doi:10.1037/0021-9010.91.1.1.

Huber, P.J., (2011). Robust Statistics, in: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1248–1251. doi:10.1007/978-3-642-04898-2_594

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235–241. doi:10.1016/j.tics.2014.02.010.

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics, 2*, 1–3.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology, 66*, 2303–2309. doi:10.1080/17470218.2013.863371.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*, 764–766. doi:10.1016/j.jesp.2013.03.013.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2015). Package e1071.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase Trust in Humans? A critical review of research. *Perspectives on Psychological Science, 10*, 772–789. doi:10.1177/1745691615600138.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422–1425. doi:10.1126/science.aab2374.

Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: a survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473–486. doi:10.1111/j.1744-6570.1991.tb02401.x.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment Research and Evaluation, 9*, 1–12.

Pollet, T. V., Cobey, K. D., & van der Meij, L. (2013). Testosterone levels are negatively associated with childlessness in males, but positively related to offspring count in fathers. *PloS One, 8*, e60018. doi:10.1371/journal.pone.0060018.

R Development Core Team (2008). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., et al. (2015). Package MASS. Retrieved from CRAN: http://cran.r-project.org/web/packages/MASS/MASS.pdf.

Ritter, G., & Gallegos, M. T. (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters, 18*, 525–539. doi:10.1016/S0167-8655(97)00049–4.

Rousseeuw, P.J., & Hubert, M., (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, 73–79. doi:10.1002/widm.2

Rousseeuw, P.J., & Leroy, A.M., (2005). Robust regression and outlier detection. John Wiley & Sons, New York.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*, 688–690. doi:10.1093/beheco/ark016.

SAS (2009). SAS/STAT® 9.2 User's Guide The SIMNORMAL Procedure - Book excerpt [WWW Document]. URL https://support.sas.com/documentation/cdl/en/statugsimnormal/61832/PDF/default/statugsimnormal.pdf (accessed 3.24.16)

Shiffler, R. E. (1988). Maximum Z scores and outliers. *The American Statistician, 42*, 79–80. doi:10.1080/00031305.1988.10475530.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. doi:10.1177/0956797611417632.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547. doi:10.1037/a0033242.

Stanton, S. J. (2011). The essential implications of gender in human behavioral endocrinology studies. *Frontiers in Behavioral Neuroscience, 5*, 9. doi:10.3389/fnbeh.2011.00009.

Stevenson, W. (2013). Simulating Random Multivariate Correlated Data (Continuous Variables) [WWW Document]. URL http://www.r-bloggers.com/simulating-random-multivariate-correlated-data-continuous-variables/ (accessed 3.23.16).

Tukey, J.W. (1977). Exploratory data analysis. Addison-Wesley, Reading, Ma.

UCLA: Statistical Consulting Group (2016). SAS Macros: corr2data [WWW Document]. URL http://www.ats.ucla.edu/stat/sas/macros/corr2data_demo.htm (accessed 3.24.16).

Van den Berg, T. (n.d.) Generating correlated random numbers [WWW Document]. URL http://www.sitmo.com/article/generating-correlated-random-numbers/#comment-325 (accessed 3.24.16).

van der Meij, L., Buunk, A. P., van de Sande, J. P., & Salvador, A. (2008). The presence of a woman increases testosterone in aggressive dominant men. *Hormones and Behavior, 54*, 640–644. doi:10.1016/j.yhbeh.2008.07.001.

van der Meij, L., Klauke, F., Moore, H. L., Ludwig, Y. S., Almela, M., & van Lange, P. A. M. (2015). Football Fan aggression: the importance of low basal cortisol and a fair referee. *PloS One, 10*, e0120103. doi:10.1371/journal.pone.0120103.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*, 1–20.

Wickham, H. (2009). ggplot2. Springer New York, New York, NY. doi:10.1007/978-0-387-98141-3

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*, 1–29. doi:10.18637/jss.v040.i01.

Wikland, K. A., Luo, Z. C., Niklasson, A., & Karlberg, J. (2002). Swedish population-based longitudinal reference values from birth to 18 years of age for height, weight and head circumference. *Acta Paediatrica, 91*, 739–754. doi:10.1080/08035250213216.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *The British Journal of Mathematical and Statistical Psychology, 57*, 173–181. doi:10.1348/000711004849222.