

Violating the normality assumption may be the lesser of two evils

Ulrich Knief^{1,*} & Wolfgang Forstmeier²

¹ Division of Evolutionary Biology, Faculty of Biology, Ludwig Maximilian University of Munich, 82152 Planegg-Martinsried, Germany

² Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany

* Address for correspondence: Ulrich Knief, Division of Evolutionary Biology, Faculty of Biology, Ludwig Maximilian University of Munich, Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany, Phone: 0049-89-2180-74101, Fax: 0049-89-2180-74104, E-mail: knief@biologie.uni-muenchen.de

Tweetable abstract: Gaussian models are remarkably robust to even dramatic violations of the normality assumption.

Key-words: hypothesis testing, linear model, normality, regression

Running header: Consequences of violating the normality assumption

Word count: ca 7700

Figures & Tables: 3 figures, 2 tables

References: 68

Supplementary Material: methods, 7 figures, 5 tables

Data availability: all scripts bundled in the R package “TrustGauss”

1 **Abstract**

2 **1.** When data are not normally distributed (e.g. skewed, zero-inflated, binomial, or count data)

3 researchers are often uncertain whether it may be legitimate to use tests that assume Gaussian errors

4 (e.g. regression, *t*-test, ANOVA, Gaussian mixed models), or whether one has to either model a more

5 specific error structure or use randomization techniques.

6 **2.** Here we use Monte Carlo simulations to explore the pros and cons of fitting Gaussian models to

7 non-normal data in terms of risk of type I error, power and utility for parameter estimation.

8 **3.** We find that Gaussian models are remarkably robust to non-normality over a wide range of

9 conditions, meaning that *P*-values remain fairly reliable except for data with influential outliers

10 judged at strict alpha levels. Gaussian models also perform well in terms of power and they can be

11 useful for parameter estimation but usually not for extrapolation. Transformation of data before

12 analysis is often advisable and visual inspection for outliers and heteroscedasticity is important for

13 assessment. In strong contrast, some non-Gaussian models and randomization techniques bear a range

14 of risks that are often insufficiently known. High rates of false-positive conclusions can arise for

15 instance when overdispersion in count data is not controlled appropriately or when randomization

16 procedures ignore existing non-independencies in the data.

17 **4.** Overall, we argue that violating the normality assumption bears risks that are limited and

18 manageable, while several more sophisticated approaches are relatively error prone and difficult to

19 check during peer review. Hence, as long as scientists and reviewers are not fully aware of the risks,

20 science might benefit from preferentially trusting Gaussian mixed models in which random effects

21 account for non-independencies in the data in a transparent way.

22 **Introduction**

23 In the biological, medical and social sciences, the validity or importance of research findings is
24 generally assessed via statistical significance tests. Significance tests ensure the trustworthiness of
25 scientific results and should reduce the amount of random noise entering the scientific literature.
26 Brunner and Austin (2009) even regard this as the “primary function of statistical hypothesis testing in
27 the discourse of science”. However, the validity of parametric significance tests may depend on
28 whether model assumptions are violated (Gelman & Hill 2007; Zuur *et al.* 2009). In a growing body
29 of literature, researchers express their concerns about irreproducible results (Open Science
30 Collaboration 2015; Ebersole *et al.* 2016; Camerer *et al.* 2018; Silberzahn *et al.* 2018) and it has been
31 argued that the inappropriate use of statistics is a leading cause of irreproducible results (Forstmeier,
32 Wagenmakers & Parker 2017). Yet researchers may often be uncertain about which statistical
33 practices enable them to answer their scientific questions effectively and which might be regarded as
34 error prone.

35
36 One of the most widely known assumptions of parametric statistics is the assumption that errors
37 (model residuals) are normally distributed (Lumley *et al.* 2002). This “normality assumption”
38 underlies the most commonly used tests for statistical significance, that is linear models “lm” and
39 linear mixed models “lmm” with Gaussian error, which includes the often more widely known
40 techniques of regression, *t*-test and ANOVA. However, empirical data often deviates considerably
41 from normality, and may even be categorical such as binomial or count data. Recent advances in
42 statistical modelling appear to have solved this problem, because it is now possible to fit generalized
43 linear mixed models “glmm” with a variety of error distributions (e.g. binomial, Poisson, zero-inflated
44 Poisson, negative binomial; O’Hara 2009; Harrison *et al.* 2018) or to use a range of randomization
45 techniques such as bootstrapping (Good 2005) in order to obtain *P*-values and confidence intervals for
46 parameter estimates from data that does not comply with any of those distributions.

47
48 While these developments have supplied experts in statistical modelling with a rich and flexible
49 toolbox, we here argue that these new tools also have created substantial damage, because they come

50 with a range of pitfalls that are often not sufficiently understood by a large majority of scientists who
51 are not outspoken experts in statistics, but who nevertheless implement the tools. The diversity of
52 possible mistakes is so large and sometimes specific to certain software applications that we only
53 want to provide some examples that we have repeatedly come across (see Box 1). Our examples
54 include failure to account for overdispersion in glmms with Poisson errors (Harrison 2014; Ives 2015;
55 Forstmeier, Wagenmakers & Parker 2017), inadequate resampling in bootstrapping techniques (e.g.
56 Ihle *et al.* 2019; Santema, Schlicht & Kempnaers 2019), as well as problems with pseudoreplication
57 due to issues with model convergence (Barr *et al.* 2013; Forstmeier, Wagenmakers & Parker 2017;
58 Arnqvist 2020). These issues may lead to anticonservative P -values and hence a high risk of false
59 positive claims.

60

61 In light of these difficulties we here want to argue whether it may often be “the lesser of two evils”
62 when researchers fit conventional Gaussian (mixed) models to non-normal data, because, as we will
63 show, Gaussian models are remarkably robust to non-normality, ensuring that type I errors (false-
64 positive conclusion) are kept at the desired low rate. Hence, we argue that for the key purpose of
65 limiting type I errors it may often be fully legitimate to model binomial or count data in Gaussian
66 models, and we also would like to raise awareness of some of the pitfalls inherent to non-Gaussian
67 models.

68

69 **Box 1** | Examples of specialized techniques that may result in biased parameter estimates or in a high
70 rate of false-positive findings due to unrecognized problems of pseudoreplication.

71 (A) Many researchers, being concerned about fitting an “inappropriate” Gaussian model, hold the
72 believe that binomial data always requires modelling a binomial error structure, and that count
73 data mandates modelling a Poisson-like process. Yet, what they consider to be “more appropriate
74 for the data at hand” may often fail to acknowledge the non-independence of events in count data
75 (Harrison 2014; Harrison 2015; Ives 2015; Forstmeier, Wagenmakers & Parker 2017). For
76 instance, in a study of butterflies choosing between two species of host plants for egg laying, an
77 individual butterfly may first sit down on species A and deposit a clutch of 50 eggs, followed by

78 a second landing on species B where another 50 eggs are laid. If we characterize the host
79 preference for species A of this individual by the total number of eggs deposited ($p(A) = 0.5$, $N =$
80 100) we obtain a highly anticonservative estimate of uncertainty (95% CI for $p(A)$: 0.398–0.602),
81 while if we base our preference estimate on the number of landings ($p(A) = 0.5$, $N = 2$) we obtain
82 a much more appropriate confidence interval (95% CI for $p(A)$: 0.013–0.987). Even some
83 methodological “how-to” guides (e.g. Fordyce *et al.* 2011; Ramsey & Schafer 2013; Harrison *et*
84 *al.* 2018) forgot to clearly explain that it is absolutely essential to model the non-independence of
85 events via random effects or overdispersion parameters (Zuur *et al.* 2009; Harrison 2014;
86 Harrison 2015; Ives 2015). Unfortunately, non-Gaussian models with multiple random effects
87 often fail to reach model convergence (e.g. Brooks *et al.* 2017), which often lets researchers
88 settle for a model that ignores non-independence and yields estimates with inappropriately high
89 confidence and statistical significance (Barr *et al.* 2013; Forstmeier, Wagenmakers & Parker
90 2017; Arnqvist 2020).

91 (B) When observational data do not comply with any distributional assumption, randomization
92 techniques like bootstrapping seem to offer an ideal solution for working out the rate at which a
93 certain estimate arises by chance alone (Good 2005). However, such resampling can also be risky
94 in terms of producing false-positive findings if the data is structured (temporal autocorrelation,
95 random effects; e.g. Ihle *et al.* 2019) and if this structure is not accounted for in the resampling
96 regime (blockwise bootstrap; e.g. Önöz & Bayazit 2012). Specifically, there is the risk that non-
97 independence introduces a strong pattern in the observed data, but, in the simulated data,
98 comparably strong patterns do not emerge because the confounding non-independencies were
99 broken up (Ihle *et al.* 2019). We argue that pseudoreplication is a well-known problem that has
100 been solved reasonably well within the framework of mixed models, and the consideration or
101 neglect of essential random effects can be readily judged from tables that present the model
102 output. In contrast, the issue of pseudoreplication is more easily overlooked in studies that
103 implement randomization tests, where the credibility of findings hinges on details of the
104 resampling procedure that are not understood by the majority of readers.

105 (C) When distributions of counts contain a high fraction of zeroes, many researchers think that this
106 issue can be fixed by specifying a zero-inflated model with Poisson or negative binomial error
107 structure. However, they may not be aware of the concept that underlies such models and hence
108 may not understand that such a model, depending on the distribution of the non-zero values, may
109 effectively treat all zeroes as missing values rather than as valid data. This could yield biased
110 parameter estimates in those cases where zeroes represent a valid phenotype rather than a case of
111 missing information. In contrast, such zero-inflated models are ideal when trying to separate two
112 processes, one that is responsible for the occurrence of (some of the) zeroes and one that is
113 responsible for variation in counts (possibly including some zeroes; Brooks *et al.* 2017).

114

115 **A wide range of opinions about violating the normality assumption**

116 Throughout the scientific literature, linear models are typically said to be robust to the violation of the
117 normality assumption when it comes to hypothesis testing and parameter estimation as long as outliers
118 are handled properly (Box & Watson 1962; Miller 1986; Ali & Sharma 1996; Lumley *et al.* 2002;
119 Gelman & Hill 2007; Zuur, Ieno & Elphick 2010; Ramsey & Schafer 2013; Williams, Grajales &
120 Kurkiewicz 2013; Puth, Neuhauser & Ruxton 2014; Warton *et al.* 2016), yet authors seem to differ
121 notably in their opinion on how serious we should take the issue of non-normality.

122

123 At one end of the spectrum, Gelman and Hill (2007) write “The regression assumption that is
124 generally *least* important is that the errors are normally distributed” and “Thus, in contrast to many
125 regression textbooks, we do not recommend diagnostics of the normality of regression residuals” (p.
126 46). At the other end of the spectrum, Osborne and Waters (2002) highlight four assumptions of
127 regression that researchers should *always* test, the first of which is the normality assumption. They
128 write “Non-normally distributed variables (highly skewed or kurtotic variables, or variables with
129 substantial outliers) can distort relationships and significance tests”. And since only few research
130 articles report having tested the assumptions underlying the tests presented, Osborne and Waters

131 (2002) worry that they are “forced to call into question the validity of many of these results,
132 conclusions and assertions”.

133

134 Between those two ends of the spectrum, many authors adopt a cautious attitude, and regard models
135 that violate the distributional assumptions as ranging from “risky” to “not appropriate”, hence
136 pleading for the use of transformations (e.g. Miller 1986; Bishara & Hittner 2012; Puth, Neuhauser &
137 Ruxton 2014), non-parametric statistics (e.g. Miller 1986), randomization procedures (e.g. Bishara &
138 Hittner 2012; Puth, Neuhauser & Ruxton 2014), or generalized linear models where the Gaussian
139 error structure can be changed to other error structures (e.g. Poisson, binomial, negative binomial) that
140 may better suit the nature of the data at hand (O'Hara 2009; O'Hara & Kotze 2010; Fordyce *et al.*
141 2011; Warton & Hui 2011; Szöcs & Schäfer 2015; Warton *et al.* 2016; Harrison *et al.* 2018). The
142 latter suggestion, however, may bear a much more serious risk: while Gaussian models are generally
143 accepted to be fairly robust to non-normal errors (here and in the following, we mean by “robust”
144 ensuring a reasonably low rate of type I errors), Poisson models are highly sensitive if their
145 distributional assumptions are violated (see Box 1), leading to a substantially increased risk of type I
146 errors if overdispersion remains unaccounted for (Warton & Hui 2011; Ives 2015; Szöcs & Schäfer
147 2015; Warton *et al.* 2016).

148

149 In face of this diverse literature, it is rather understandable that empirical researchers are largely
150 uncertain about the importance of adhering to the normality assumption in general, and about how
151 much deviation and which form of deviation might be tolerable under which circumstances (in terms
152 of sample size and significance level threshold). With the present article we hope to provide
153 clarification and guidance.

154

155 We here use Monte Carlo simulations to explore how violations of the normality assumption affect
156 the probability of drawing false-positive conclusions (the rate of type I errors), because these are the
157 greatest concern in the current reliability crisis (Open Science Collaboration 2015). We aim at
158 deriving simple rules of thumb, which researchers can use to judge whether the violation may be

159 tolerable and whether the P -value can be trusted. We also assess the effects of violating the normality
160 assumption in terms of bias and precision on parameter estimation. Furthermore, we provide an R
161 package (“TrustGauss”) that researchers can use to explore the effect of specific distributions on the
162 reliability of P -values and parameter estimates.

163

164 Counter to intuition, but consistent with a considerable body of literature (Box & Watson 1962; Miller
165 1986; Ali & Sharma 1996; Lumley *et al.* 2002; Gelman & Hill 2007; Zuur, Ieno & Elphick 2010;
166 Ramsey & Schafer 2013; Williams, Grajales & Kurkiewicz 2013; Puth, Neuhauser & Ruxton 2014;
167 Warton *et al.* 2016) we find that violations of the normality of residuals assumption are rarely
168 problematic for hypothesis testing and parameter estimation, and we argue that the commonly
169 recommended solutions may bear greater risks than the one to be solved.

170

171 **The linear regression model and its assumptions**

172 At this point we need to briefly introduce the notation for the model of least squares linear regression.
173 In its simplest form, it can be formulated as $Y_i = a + b \times X_i + e_i$, where each element of the dependent
174 variable Y_i is linearly related to the predictor X_i through the regression coefficient b (slope) and the
175 intercept a . e_i is the error or residual term, which describes the deviations (residuals) of the actual
176 from the true unobserved (error) or the predicted (residual) Y_i and whose sum equals zero (Sokal &
177 Rohlf 1995; Gelman & Hill 2007). An F -test is usually employed for testing the significance of
178 regression models (Ali & Sharma 1996).

179

180 Basic statistics texts introduce (about) five assumptions that need to be met for interpreting all
181 estimates from linear regression models safely (**Box 2**: validity, independence, linearity,
182 homoscedasticity of the errors and normality of the errors; Gelman & Hill 2007). Out of these
183 assumptions, normally distributed errors are generally assumed to be the least important (yet probably
184 the most widely known; Lumley *et al.* 2002; Gelman & Hill 2007). Deviations from normality usually
185 do not bias regression coefficients (Ramsey & Schafer 2013; Williams, Grajales & Kurkiewicz 2013)
186 or impair hypothesis testing (no inflated type I error rate, e.g. Bishara & Hittner 2012; Ramsey &

187 Schafer 2013; Puth, Neuhauser & Ruxton 2014; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.*
188 2016) even at relatively small sample sizes. With large sample sizes ≥ 500 the Central Limit Theorem
189 guarantees that the regression coefficients are on average normally distributed (Ali & Sharma 1996;
190 Lumley *et al.* 2002).

191

192 **Box 2** | Five assumptions of regression models: validity, independence, linearity, homoscedasticity of
193 the errors and normality of the errors (Gelman & Hill 2007). Three of these criteria are concerned
194 with the dependent variable Y , or — to be more precise — the regression error e (assumptions 2, 4 and
195 5, see below). The predictor X is often not considered, although e is supposed to be normal and of
196 equal magnitude at every value of X .

- 197 (1) *Validity* is not a mathematical assumption *per se* but it still poses “the most challenging step in
198 the analysis” (Gelman & Hill 2007), namely that regression should enable the researcher to
199 answer the scientific question at hand (Kass *et al.* 2016).
- 200 (2) Each value of the dependent variable Y is influenced by only a single value of the predictor X ,
201 meaning that all observations and regression errors e_i are *independent* (Quinn & Keough 2002).
202 Dependence among observations commonly arises either through cluster (i.e. data collected on
203 subgroups) or serial effects (i.e. data collected in temporal or spatial proximity; Ramsey &
204 Schafer 2013). We will discuss the independence assumption later because it is arguably the
205 riskiest to violate in terms of producing type I errors (Zuur *et al.* 2009; see “A word of caution”).
- 206 (3) The dependent variable Y and the predictors should be *linearly* (and additively) related through
207 the regression coefficient b . That being said, quadratic or higher-order polynomial relationships
208 can also be accommodated by squaring or raising the predictor variable X to a higher power,
209 because Y is still modelled as a linear function through the regression coefficient (Williams,
210 Grajales & Kurkiewicz 2013).
- 211 (4) The variance in the regression error e (or the spread of the response around the regression line) is
212 constant across all values of the predictor X , i.e. the samples are *homoscedastic*. Deviations from
213 homoscedasticity will not bias parameter estimates of the regression coefficient b (Gelman &
214 Hill 2007). Slight deviations are thought to have only little effects on hypothesis testing (Osborne

215 & Waters 2002) and can often be dealt with by weighted regression, mean-variance stabilizing
216 data transformations (e.g. log-transformation) or estimation of heteroscedasticity-robust standard
217 errors (Huber 1967; White 1980; Miller 1986; Zuur *et al.* 2009; see “A word of caution” for
218 further discussion).

219 (5) The errors of the model should be normally distributed (*normality* assumption), which should be
220 tested via inspecting the distribution of the model residuals e (Zuur, Ieno & Elphick 2010). Both
221 visual approaches (probability or QQ-plots) and formal statistical tests (Shapiro-Wilk) are
222 commonly applied. Formal tests for normality have been criticized because they have low power
223 at small sample sizes and almost always yield significant deviations from normality at large
224 sample sizes (Ghasemi & Zahediasl 2012). Thus, researchers are mostly left with their intuition
225 to decide how severely the normality assumption is violated and how robust regression is to such
226 violations. A researcher who examines the effect of a single treatment on multiple dependent
227 variables (e.g. health parameters) may adhere strictly to the normality assumption and thus
228 switch forth and back between reporting parametric and non-parametric test statistics depending
229 on how strongly the trait of interest deviates from normality, rendering a comparison of effect
230 sizes difficult.

231
232 Importantly, the robustness of regression methods to deviations from normality of the regression
233 errors e does not only depend on sample size, but also on the distribution of the predictor X (Box &
234 Watson 1962; Mardia 1971). Specifically, when the predictor variable X contains a single outlier, then
235 it is possible that the case coincides with an outlier in Y , creating an extreme observation with high
236 leverage on the regression line. This is the only case where statistical significance gets seriously
237 misestimated based on the assumption of Gaussian errors in Y which is violated by the outlier in Y .
238 This problem has been widely recognized (Box & Watson 1962; Miller 1986; Ali & Sharma 1996;
239 Osborne & Waters 2002; Zuur, Ieno & Elphick 2010; Ramsey & Schafer 2013) leading to the
240 conclusion that Gaussian models are robust as long as there are no outliers that occur in X and Y
241 simultaneously. Conversely, violations of the normality assumption that do not result in outliers
242 should not lead to elevated rates of type I errors.

243

244 Distributions of empirical data may deviate from a Gaussian distribution in multiple ways. Rather
245 than being continuous, data may be discrete, such as integer counts or even binomial character states
246 (yes/no data). Continuous variables may deviate from normality in terms of skewness (showing a long
247 tail on one side), kurtosis (curvature leading to light or heavy tails), and even higher-order moments.
248 All these deviations are generally thought to be of little concern (e.g. Bishara & Hittner 2012), even if
249 they are far from fitting to the bell-shaped curve, such as binomial data (Cochran 1950). However,
250 heavily skewed distributions typically result in outliers, which, depending on the distribution of X , can
251 be problematic in terms of type I error rates as just explained above (see also Blair & Lawson 1982).
252 In our simulations we try to representatively cover much of the diversity in possible distributions, in
253 order to provide a broad overview that extends beyond the existing literature. We focus on fairly
254 drastic non-normality because only little bias can be expected from minor violations (Hack 1958;
255 Glass, Peckham & Sanders 1972; Bishara & Hittner 2012; Puth, Neuhauser & Ruxton 2014).

256

257 **Simulations to assess effects on P -values, power and parameter estimates**

258 To illustrate the consequences of violating the normality assumption, we performed Monte Carlo
259 simulations on five continuous and five discrete distributions that were severely skewed, platy- and
260 leptokurtic or zero-inflated (distributions D0–D9 in **Figure 1A** left column, **Table 1**), going beyond
261 previous studies that examined less dramatic violations (Bishara & Hittner 2012; Puth, Neuhauser &
262 Ruxton 2014; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016) but that are still of biological
263 relevance (Gelman & Hill 2007; Frank 2009; Zuur *et al.* 2009). For example, measures of fluctuating
264 asymmetry are distributed half-normally (distribution D4, **Table 1**) or survival data can be modelled
265 using a gamma distribution (distribution D9, **Table 1**). The R-code for generating these distributions
266 can be found in the R package “TrustGauss” in the Supplementary Material, where we also provide
267 the specific parameter settings used for generating distributions D0–D9. Moments of these
268 distributions are provided in **Table 1**. We explored these 10 distributions across a range of sample
269 sizes ($N = 10, 25, 50, 100, 250, 500, 1000$). Starting with the normal distribution D0 for reference, we
270 sorted the remaining distributions D1–D9 by increasing tendency to produce strong outliers because

271 these are known to be problematic (calculated as the average proportion of data points with Cook's
272 distance exceeding a critical value (see below) at a sample size of $N = 10$). We used these data both as
273 our dependent variable Y and as our predictor variable X in linear regression models, yielding 10×10
274 = 100 combinations of Y and X for each sample size (see **Figure S1** for distributions of the
275 independent variable Y , the predictor X , and residuals). A detailed documentation of the TrustGauss-
276 functions and their application is provided in the Supplement.

277

278 We assessed the significance of the models via an F -test wherever possible and used a likelihood ratio
279 test otherwise. We fitted these models to 50,000 datasets for each combination of the dependent and
280 predictor variable. We did not simulate any effect, which means that both the regression coefficient b
281 and the intercept a were on average zero. This enabled us to use the frequency of all models that
282 yielded a P -value ≤ 0.05 as an estimate of the type I error rate at a significance level (α) of 0.05. The
283 null distribution of P -values is uniform on the interval $[0,1]$ and because all P -values are independent
284 and identically distributed, we constructed concentration bands using a beta-distribution (cf. Casella
285 & Berger 2002; Knief *et al.* 2017; QQ-plots of expected vs observed P -values are depicted in **Figure**
286 **S1**). We assessed the deviation of observed from expected $-\log_{10}(P$ -values) at an expected exponent
287 value of 3 ($P = 10^{-3}$; $-\log_{10}(10^{-3}) = 3$) and 4 ($P = 10^{-4}$) and by estimating the scale shift parameter $\nu =$
288 $\sigma_{\text{observed}} / \sigma_{\text{expected}}$ (Lin 1989), where σ is the standard deviation in $-\log_{10}(P$ -values). We further
289 calculated studentized residuals (R), hat values (H) and Cook's distances (D) as measures of
290 discrepancy, leverage and influence, respectively, and assessed which proportion exceeded critical
291 values of $R > 2$, $H > (2 \times (k + 1)) / n$ and $D > 4 / (n - k - 1)$, where k is the number of regression slopes
292 and n is the number of observations (Zuur, Ieno & Smith 2007).

293

294 Since some of the predictor variables were binary rather than continuous, our regression models also
295 comprise the situation of classical two-sample t -tests, and we assume that the results would also
296 generalize to the situation of multiple predictor levels (ANOVA), which can be decomposed to
297 multiple binary predictors. To demonstrate that our conclusions from univariate models (involving a
298 single predictor) generalize to the multivariate case (involving several predictors), we fitted the above

299 models with a sample size of $N = 100$ to the same 10 dependent variables with three normally
300 distributed predictors and one additional predictor sampled from the 10 different distributions. We
301 further fitted the above models as mixed-effects models using the lme4 R package (v1.1-14, Bates *et*
302 *al.* 2015). For that we simulated $N = 100$ independent samples each of which was sampled twice, such
303 that the single random effect “sample ID” explained roughly 30% of the variation in Y . We encourage
304 readers to try their own simulations using our R package “TrustGauss”.

305

306 We evaluated power, bias and precision of parameter estimates using a sample size of $N = 10, 100,$
307 1000 and the same 10 distributions of the independent and dependent variables as above. We
308 simulated multivariate data by first Z-transforming the independent variable Y and the covariate X .
309 We then used an iterative algorithm (SI technique, Ruscio & Kaczetow 2008) that samples from the
310 Z-transformed distributions of Y and X to introduce a predefined effect size of $r = 0.15, 0.2$ and 0.25
311 in 50,000 simulations. Additionally, to remove the dominating effect of sample size on power
312 calculations, we calculated the effect size that would be needed to reach a power of 0.5 (rounded to
313 the third decimal) for $N = 10, 100$ and 1000 if Y and X were normally distributed using the
314 powerMediation R package (v0.2.9, Dupont & Plummer 1998; Qiu 2018). This yielded effect sizes of
315 0.59, 0.19 and 0.062, respectively. We then introduced effects of such magnitudes with their
316 respective sample sizes in 50,000 simulations. For distribution D6 and the combinations of D8 with
317 D9 we were unable to introduce the predefined effect size also at very large sample sizes ($N =$
318 $100,000$) and we removed those from further analyses. We estimated power (β) as the proportion of
319 all simulations that yielded a significant (at $\alpha = 0.05$ or $\alpha = 0.001$) regression coefficient b . In the case
320 of normally distributed Y and X , this yielded power estimates that corresponded well with the
321 expectations calculated using the powerMediation R package (v0.2.9, **Table S1**, Dupont & Plummer
322 1998; Qiu 2018). We used the mean and the coefficient of variation (CV) of the regression coefficient
323 b as our measures of bias and precision, respectively. We also assessed interpretability and power of
324 Gaussian versus binomial (mean = 0.75) and Poisson (mean = 1) at a sample size of $N = 100$ by fitting
325 models with a Gaussian, binomial or Poisson error structure in the glms. The effect sizes were chosen

326 such that we reached a power of around 0.5 (see **Table S2** for details on distributions and effect sizes)
327 and models were fitted to 50,000 of such datasets.

328

329 **Results**

330 *Effects on P-values*

331 The rate at which linear regression models with Gaussian error structure produced false-positive
332 results (type I errors) was very close to the expected value of 0.05 (**Figure 1B**). When sample size
333 was high ($N = 1000$), type I error rates ranged only between 0.044 and 0.052, across the 100
334 combinations of distributions of the dependent variable Y and the predictor X . Hence, despite of even
335 the most dramatic violations of the normality assumption (see e.g. distributions D8 and D9 in **Figure**
336 **1A**), there was no increased risk of obtaining false-positive results. At $N = 100$, the range was still
337 remarkably narrow (0.037–0.058), and only for very low sample sizes ($N = 10$) we observed 4 out of
338 100 combinations which yielded notably elevated type I error rates in the range of 0.086 to 0.11.
339 These four cases all involved combinations of the distributions D8 and D9, which yield extreme
340 leverage observations (**Figure S2**). For this low sample size of $N = 10$, there were also cases where
341 type I error rates were clearly too low (down to 0.015, involving distributions D1–D3 where extreme
342 values are rarer than under the normal distribution D0; for details see **Figure S2** and **Table S3**).

343

344 Next, we examine the scale shift parameter (**Figure 1C**) which evaluates the match between observed
345 and expected distributions of P -values across the entire range of P -values (not only the fraction at the
346 5% cut-off). Whenever either the dependent variable Y or the predictor X was normally distributed,
347 the observed and expected P -values corresponded very well (first row and first column in **Figure 1C**).
348 Accordingly, the P -values fell within the 95% concentration bands across their entire range (rightmost
349 column in **Figures S1**). This observation was unaffected by sample size (**Table S4**). However, if both
350 the dependent variable Y and the predictor X were heavily skewed, consistently inflated P -values
351 outside the concentration bands occurred, yet this was almost exclusively limited to the case of $N = 10$
352 (**Figure 1C**). For larger sample sizes only the most extreme distribution D9 produced somewhat
353 unreliable P -values (**Figure 1C**). This latter effect of unreliable (mostly anti-conservative) P -values

354 was most pronounced when judgements were made at a very strict α -level (**Figure 1D** $\alpha = 0.001$ and
355 **Figure 1E** $\alpha = 0.0001$). At a sample size of $N = 100$, and for $\alpha = 0.001$, observed $-\log_{10}(P\text{-values})$
356 were biased maximally 3.36-fold when both X and Y were sampled from distribution D9. This means
357 that P -values of about $P = 10^{-10}$ occurred at a rate of 0.001 ($P = 10^{-(3 \times 3.36)} = 10^{-10.08}$; **Figure 1D**). At N
358 $= 100$, and for $\alpha = 0.0001$, the bias was maximally 4.54-fold (**Figure 1E**). Our multivariate and
359 mixed-model simulations confirmed that these patterns are general and also apply to models with
360 multiple predictor variables (**Figure S3**) and to models with a single random intercept (**Figure S4**).
361
362 Based on the 100 simulated scenarios that we have constructed, P -values from Gaussian models are
363 highly robust to even extreme violation of the normality assumption and can be trusted, except when
364 involving X and Y distributions with extreme outliers (distribution D9; see also Blair & Lawson
365 1982). For very small sample sizes, judgements should preferably be made at $\alpha = 0.05$ (rather than at
366 more strict thresholds) and should also beware of outliers in both X and Y . The same distributions of
367 the dependent and the independent variable introduced the same type I error rates, meaning that
368 effects were symmetric (Box & Watson 1962). We reference the reader to the “A word of caution”
369 section, where we discuss both the assumption of equal variances of the errors and the effects of non-
370 normality on other applications of linear regression.

371

372 *Effects on power and parameter estimates*

373 Power of linear regression models with a Gaussian error structure was only weakly affected by the
374 distributions of Y and X , whereas sample size and effect size were much more influential (**Figure 2B**,
375 **Figures S5B, S6B**). Power appears to vary notably between distributions when sample size and hence
376 power are small ($N = 10$ in **Figure 2B**), but this variability rather closely reflects the corresponding
377 type I error rates shown in **Figure 1B** (Pearson correlation $r = 0.69$ between **Figure 1B** and **2B** across
378 the $N = 79$ combinations with power estimates at regression coefficient $b = 0.2$ and sample size $N =$
379 10). To assess the effects of sample size and non-normality on power, we adjusted the regression
380 coefficients such that power stayed constant at 50% for normally distributed Y and X at sample sizes
381 of $N = 10, 100$ and 1000 ($b = 0.59, 0.19$ and 0.062 , respectively, **Figure 2C**). Then, for $N = 1000$,

382 power was essentially unaffected by the distribution of Y and X , ranging from 0.48 to 0.52 for all but
383 one combination of Y and X ($\beta = 0.45$ when Y and X are distributed as D9, that is gamma $\Gamma(0.1, 100)$,
384 **Table 1**). In that particular combination, power was not generally reduced but the distribution of P -
385 values was shifted, such that power could either be reduced or increased depending on the α -threshold
386 (at $\alpha = 0.001$ that combination yielded the highest power). At $N = 100$, power varied slightly more
387 (0.44–0.60) but still 87% of all power estimates were between 0.48 and 0.52. Only at a sample size of
388 $N = 10$, power varied considerably between 0.05 and 0.87 (30% of all estimates between 0.48 and
389 0.52, **Figure 2C**).

390

391 For most distributions of Y and X , regression coefficients were unbiased, which follows from the
392 Lindeberg-Feller Central Limit Theorem (Lumley *et al.* 2002). The strongest bias occurred at a
393 sample size of $N = 10$ and when the distribution of X was highly skewed (D9), resulting in such a high
394 frequency of high leverage observations that the Lindeberg-Feller Central Limit Theorem did not hold
395 (**Figure S2**). In the most extreme case, the mean regression coefficients at $N = 10$ were below zero
396 (indicated as additional white squares in **Figure S5D, S6D**). However, the bias shrunk to maximally
397 1.32-fold when the sample size increased to $N = 100$ and to 1.03-fold at a sample size of $N = 1000$
398 (**Figure 2D**).

399

400 We used the coefficient of variation in regression coefficients as our measure of the precision of
401 parameter estimates. Similar to the pattern in bias, regression coefficients were precise for most
402 distributions of Y and X and the lowest precision occurred at a sample size of $N = 10$ and when the
403 distribution of X was highly skewed (D9). However, there was no gain in precision when increasing
404 the sample size from $N = 100$ to $N = 1000$ (**Figure 2E**) and precision slightly decreased at larger
405 effect sizes (**Figure S5E, S6E**).

406

407 We conclude that in our 79 simulated scenarios neither power nor bias or precision of parameter
408 estimates are heavily affected by violations of the normality assumption by both the distributions of
409 the dependent variable Y and the predictor X , except when involving predictors with extreme outliers

410 (i.e. high leverage, distribution D9). An increase in sample size protects against severely biased
411 parameter estimates but does not make estimates more precise. We provide further advice in the “A
412 word of caution” section.

413

414 *Comparison between error distributions*

415 In the previous section, we have shown that Gaussian models are robust to violations of the normality
416 assumption. How do they perform in comparison to Poisson and binomial models and how do Poisson
417 models perform if their distributional assumptions are violated? To address these questions, we fitted
418 glms with a Gaussian, Poisson or binomial error structure to data where the dependent variable Y was
419 Gaussian, Poisson or binomial distributed and the predictor variable X followed a Gaussian, gamma or
420 binomial distribution. This allowed us to directly compare the effect of the error structure on power,
421 bias and precision of the parameter estimate. Interestingly, models with a Gaussian error structure
422 were largely comparable in terms of power and bias to those fitted using the appropriate error
423 structure. However, parameter estimates were less precise using the Gaussian error structure (**Table**
424 **2**), which argues in favour of the more specialized models for the purpose of parameter estimation.

425

426 More importantly for the reliability of science, and in contrast to Gaussian models, Poisson models
427 are not at all robust to violations of the distribution assumption. For comparison, we fitted the above
428 univariate models involving the five discrete distributions (D1, D2, D6, D7, D8) with a sample size of
429 $N = 100$ using a Poisson error structure (inappropriately). This yielded heavily biased type I error rates
430 (at $\alpha = 0.05$) in either direction ranging from 0 to as high as 0.55 (**Figure 3**, right column, **Figures**
431 **S7**). Yet when also inappropriately modelling these distributions as Gaussian, type I error rates are
432 very close to the nominal level of 0.05 (**Figure 3**, left column). Controlling for overdispersion in
433 counts through the use of a glmm with an observation-level random effect (Harrison *et al.* 2018) fixed
434 the problem of inflated type I error rates for distributions D2 and D7 (**Figure 3**, indicated in red) but
435 did not solve the problem of low power for distributions D1, D6, and D8 (**Figure 3**, indicated in blue).
436 Using a quasi-likelihood method (“Quasipoisson”, Wedderburn 1974) provided unbiased type I error

437 rates, like in the Gaussian models (**Figure 3**), but this quasi-likelihood method is not available in the
438 mixed-effects package lme4 in R (Bates *et al.* 2015).

439

440 **A word of caution**

441 Our finding that violations of the normality assumption are relatively unproblematic with regard to
442 type I errors should not be misunderstood as a *carte blanche* to violate any assumption of linear
443 models. The probably riskiest assumption to violate (in terms of producing type I errors) is the
444 assumption of independence of data points (Kass *et al.* 2016; Forstmeier, Wagenmakers & Parker
445 2017), because one tends to overestimate the amount of independent evidence that is provided by the
446 data points, which are not real replicates (hence this is called “pseudoreplication”).

447

448 Another assumption that is not to be ignored concerns the homogeneity of variances across the entire
449 range of the predictor variable (Box 1953; Glass, Peckham & Sanders 1972; Miller 1986;
450 McGuinness 2002; Osborne & Waters 2002; Zuur *et al.* 2009; Ramsey & Schafer 2013; Williams,
451 Grajales & Kurkiewicz 2013). Violating this assumption may result in more notable increases of type
452 I errors (compared to what we examined here) at least when the violations are drastic. For instance,
453 when applying a *t*-test that assumes equal variances in both groups to data that come from
454 substantially different variances (e.g. $\sigma_1^2/\sigma_2^2 = 0.1$), then high rates of type I errors (e.g. 23%) may be
455 obtained in a situation where sample sizes are unbalanced ($N_1 = 15$, $N_2 = 5$), namely when the small
456 sample comes from the more variable group (Glass, Peckham & Sanders 1972; Miller 1986). Also in
457 this example, it is the influence of outliers (small *N* sampled from large variance) that results in
458 misleading *P*-values. We further carried out some extra simulations to explore whether non-normality
459 tends to exacerbate the effects of heteroscedasticity on type I error rates, but we found that normal and
460 non-normal data behaved practically in the same way (see Supplementary Methods and **Table S5**).
461 Hence, heteroscedasticity can be problematic, but this seems to be fairly independent of the
462 distribution of the variables.

463

464 Diagnostic plots of model residuals over fitted values can help identifying outliers and recognizing
465 heterogeneity in variances over fitted values. Transformation of variables is often a helpful remedy if
466 one observes that variance strongly increases with the mean. This typically occurs in comparative
467 studies, where e.g. body size of species may span several orders of magnitude (calling for a log-log
468 plot). Most elegantly, heteroscedasticity can be modelled directly, for instance by using the “weights”
469 argument in `lme` (see Pinheiro & Bates 2000, p. 214), which also enables us to test directly whether
470 allowing for heteroscedasticity increases the fit of the model significantly. Similarly,
471 heteroscedasticity-consistent standard errors could be estimated (Hayes & Cai 2007). For more advice
472 on handling heteroscedasticity see McGuinness (2002).

473

474 Another word of caution when running Gaussian models on non-Gaussian data should be expressed
475 when it comes to the interpretation of parameter estimates of models. If the goal of modelling lies in
476 the estimation of parameters (rather than hypothesis testing) then such models should be regarded
477 with caution. First, recall that distributions with extreme outliers are often better characterized by their
478 median than by their mean, which gets pulled away by extreme values. Second, parameter estimates
479 for counts or binomial traits may be acceptable for interpretation when they refer to the average
480 condition (e.g. a typical family having 1.8 children consisting of 50% boys). However, parameter
481 estimates may become nonsensical outside the typical range of data (e.g. negative counts or
482 probabilities). In such cases one might also consider fitting separate models for parameter estimation
483 and for hypothesis testing (Warton *et al.* 2016).

484

485 Finally, in the above we were exclusively concerned with associations between variables, that is
486 parameter estimates derived from the whole population of data points. However, sometimes we might
487 be interested in predicting the response of specific individuals in the population and we need to
488 estimate a prediction interval. In that case, a valid prediction interval requires the normality
489 assumption to be fulfilled because it is based directly on the distribution of Y (Lumley *et al.* 2002;
490 Ramsey & Schafer 2013).

491

492 **The issue of overdispersion in non-Gaussian models**

493 We have shown that Poisson models yielded heavily biased type I error rates (at $\alpha = 0.05$) in either
494 direction ranging from 0 to as high as 0.55 when their distribution assumption is violated (**Figure 3**
495 right column, **Figures S7**). This of course is an inappropriate use of the Poisson model, but still this is
496 not uncommonly found in the scientific literature. Such inflations of type I error rates in glms already
497 have been reported frequently (Young, Campbell & Capuano 1999; Warton & Hui 2011; Ives 2015;
498 Szöcs & Schäfer 2015; Warton *et al.* 2016) and this problem threatens the reliability of research
499 whenever such models are implemented with insufficient statistical expertise.

500

501 First, it is absolutely essential to control for overdispersion in the data (that is more extreme counts
502 than expected under a Poisson process), either by using a quasi-likelihood method (“Quasipoisson”)
503 or by fitting an observation level random effect (“OLRE”; **Figure 3**). Overdispersion may already be
504 present when counts refer to discrete natural entities (for example counts of animals), but may be
505 particularly strong when Poisson errors are less appropriately applied to measurements of areas (e.g.
506 counts of pixels or mm²), latencies (e.g. counts of seconds), or concentrations (e.g. counts of
507 molecules). Similarly, there may also be overdispersion in counts of successes versus failures that are
508 being analysed in a binomial model (e.g. fertile versus infertile eggs within a clutch). Failure to
509 account for overdispersion (as in **Figure 3B** and **3D**) will typically result in very high rates of type I
510 errors (Young, Campbell & Capuano 1999; Warton & Hui 2011; Ives 2015; Szöcs & Schäfer 2015;
511 Warton *et al.* 2016; Forstmeier, Wagenmakers & Parker 2017).

512

513 Second, even after accounting for overdispersion, some models may still yield inflated or deflated
514 type I error rates (not observed in our examples of **Figure 3**), therefore requiring statistical testing via
515 a resampling procedure (Warton & Hui 2011; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016),
516 but this may also depend on the software used. While several statistical experts have explicitly
517 advocated for such a sophisticated approach to count data (O'Hara 2009; O'Hara & Kotze 2010; Szöcs
518 & Schäfer 2015; Warton *et al.* 2016; Harrison *et al.* 2018), we are concerned about practicability
519 when non-experts have to make decisions about the most adequate resampling procedure, particularly

520 when there are also non-independencies in the data (random effects) that have to be considered. In this
521 field of still developing statistical approaches it seems much easier to get things wrong (and obtain a
522 highly overconfident P -value) than to get everything right (Bolker *et al.* 2009).

523

524 In summary, we are worried that authors being under pressure to present statistically significant
525 findings will misinterpret type I errors (due to incorrect implementation) optimistically as a true
526 finding and misattribute the gained significance to a presumed gain of power when fitting the
527 “appropriate” error structure (note that such power gains should be quite small; see **Table 2** and also
528 Szöcs & Schäfer 2015; Warton *et al.* 2016). Moreover, we worry that sophisticated methods may
529 allow presenting nearly anything as statistically significant (Simmons, Nelson & Simonsohn 2011)
530 because complex methods will only rarely be questioned by reviewers.

531

532 **Practical advice**

533 Anti-conservative P -values usually do not arise from violating normality in Gaussian models (except
534 for the case of influential outliers), but rather from various kinds of non-independencies in the data
535 (see Box 1). We therefore recommend the Gaussian mixed-effect model as a trustworthy and
536 universal standard tool for hypothesis testing, where transparent reporting of the model’s random
537 effect structure clarifies to the reader which non-independencies in the data were accounted for. Non-
538 normality should not be a strong reason for switching to a more specialized technique, at least not for
539 hypothesis testing, and such techniques should only be used with a good understanding of the risks
540 involved (see Box 1).

541

542 To avoid the negative consequences of strong deviations from normality that may occur under some
543 conditions (see **Figure 1**) it may be most advisable to apply a rank-based inverse normal (RIN)
544 transformation (aka rankit scores, Bliss 1967) to the data, which can approximately normalize most
545 distributional shapes and which effectively minimizes type I errors and maximises statistical power
546 (Bishara & Hittner 2012; Puth, Neuhauser & Ruxton 2014). Note that we have avoided
547 transformations in our study simply to explore the consequences of major non-normality, but we agree

548 with the general wisdom that transformations can mitigate problems with outliers (Osborne &
549 Overbay 2004), heteroscedasticity (McGuinness 2002), and sometimes with interpretability of
550 parameter estimates.

551

552 In practice, we recommend the following to referees:

553 (1) When a test assumes Gaussian errors, request a check for influential observations, particularly if
554 very small P -values are reported. Consider recommending a RIN-transformation or other
555 transformations for strong deviations from normality.

556 (2) For Poisson models or binomial models of counts, always check whether the issues of
557 overdispersion and resampling are addressed, otherwise request an adequate control for type I errors
558 or verification with Gaussian models.

559 (3) For randomization tests, request clarity about whether observed patterns may be influenced by
560 non-independencies in the data that are broken up by the randomization procedure. If so, ask for
561 possible alternative ways of testing or of randomizing (e.g. blockwise bootstrap).

562 (4) When requesting a switch to more demanding techniques (e.g. non-Gaussian models,
563 randomization techniques), reviewers should accompany this recommendation with sufficient advice,
564 caveats and guidance to ensure a safe and robust implementation.

565

566 **Conclusion**

567 If we are interested in statistical hypothesis testing, linear regression models with a Gaussian error
568 structure are generally robust to violations of the normality assumption. Judging P -values at the
569 threshold of $\alpha = 0.05$ is nearly always safe, but if both Y and X are skewed, we should avoid being
570 overly confident in very small P -values and examine whether these result from outliers in both X and
571 Y (see also Blair & Lawson 1982; Osborne & Overbay 2004). With this caveat in mind, violating the
572 normality assumption is relatively unproblematic and there is much to be gained when researchers
573 follow a standardized way of reporting effect sizes (Lumley *et al.* 2002). This is good news also for
574 those who want to apply models with Gaussian error structure to binomial or count data when models
575 with other structures fail to reach convergence or produce nonsensical estimates (e.g. Ives & Garland

576 2014; Plaschke *et al.* 2019). While Gaussian models are rarely misleading, other approaches (see
577 examples in Box 1) may bear a non-trivial risk of yielding anti-conservative *P*-values when applied by
578 scientists with limited statistical expertise.

579

580 **Data availability**

581 All functions are bundled in an R package named “TrustGauss”. The R package, R scripts,
582 supplementary figures S1, S3, S4 and S7 and the raw simulation outputs are accessible through the
583 Open Science Framework (https://osf.io/r5ym4/?view_only=5d79da4f8b4441e1addf99b0d435a45e).

584

585 **Acknowledgements**

586 We thank N. Altman, S. Nakagawa, M. Neuhäuser, F. Korner-Nievergelt and H. Schielzeth for
587 helpful discussions and B. Kempnaers and J.B.W. Wolf for their support.

588

589 **Author contributions**

590 W.F. and U.K. conceived of the study. U.K. wrote the simulation code. U.K. and W.F. prepared the
591 manuscript.

592

593 **Competing interests**

594 The authors declare no competing financial interests.

595

596 **References**

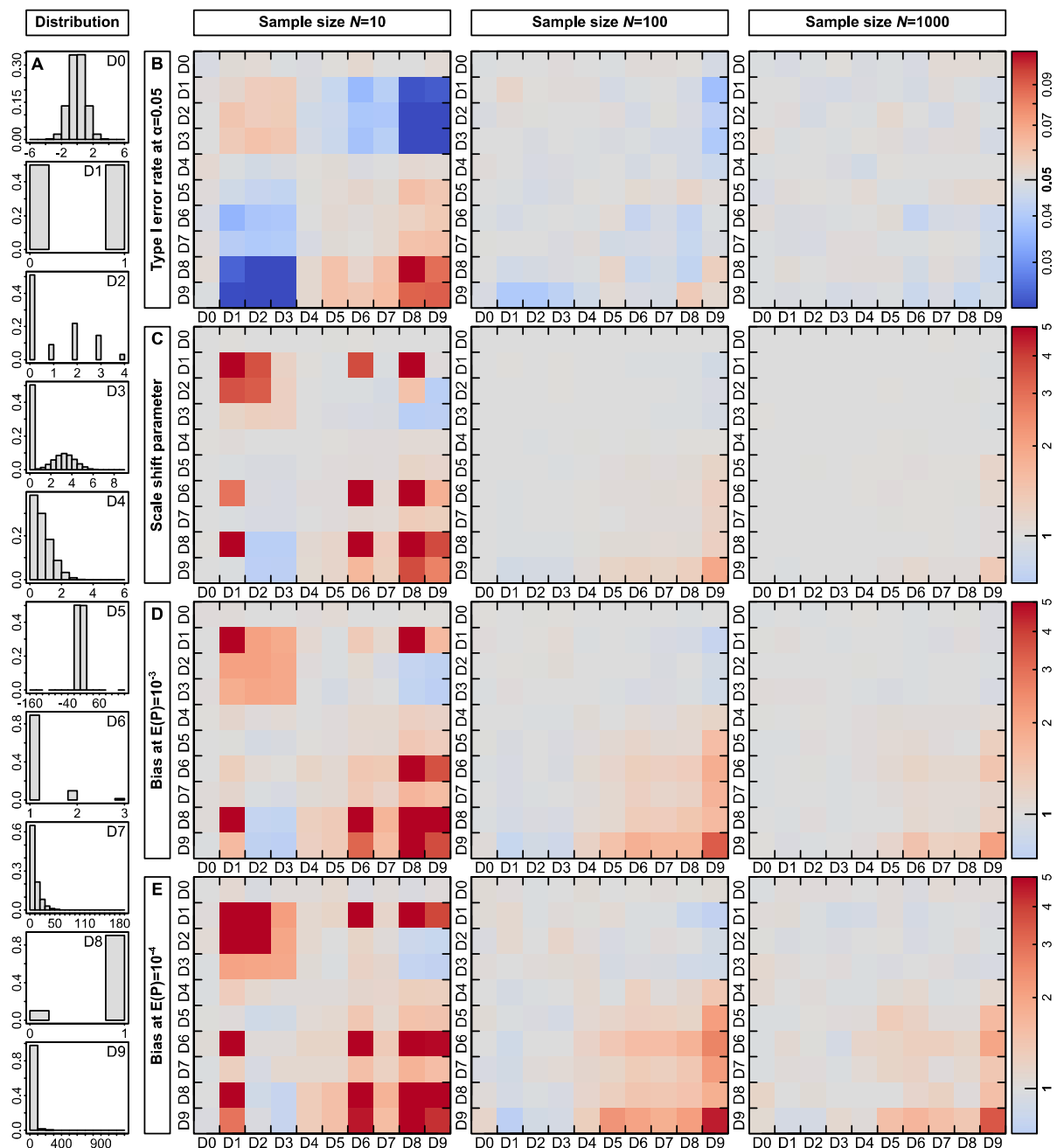
- 597 Ali, M.M. & Sharma, S.C. (1996) Robustness to nonnormality of regression F-tests. *Journal of*
598 *Econometrics*, **71**, 175–205.
- 599 Arnqvist, G. (2020) Mixed models offer no freedom from degrees of freedom. *Trends in Ecology &*
600 *Evolution*.
- 601 Barr, D.J., Levy, R., Scheepers, C. & Tily, H.J. (2013) Random effects structure for confirmatory
602 hypothesis testing: keep it maximal. *Journal of Memory and Language*, **68**, 255–278.
- 603 Bates, D., Mächler, M., Bolker, B.M. & Walker, S.C. (2015) Fitting linear mixed-effects models
604 using lme4. *Journal of Statistical Software*, **67**, 1–48.
- 605 Bishara, A.J. & Hittner, J.B. (2012) Testing the significance of a correlation with nonnormal data:
606 comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological*
607 *Methods*, **17**, 399–417.
- 608 Blair, R.C. & Lawson, S.B. (1982) Another look at the robustness of the product-moment correlation
609 coefficient to population non-normality. *Florida Journal of Educational Research*, **24**, 11–15.

- 610 Bliss, C.I. (1967) *Statistics in biology*. McGraw-Hill, New York, NY.
- 611 Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White,
612 J.S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution.
613 *Trends in Ecology & Evolution*, **24**, 127–135.
- 614 Box, G.E.P. (1953) Non-normality and tests on variances. *Biometrika*, **40**, 318–335.
- 615 Box, G.E.P. & Watson, G.S. (1962) Robustness to non-normality of regression tests. *Biometrika*, **49**,
616 93–106.
- 617 Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug,
618 H.J., Mächler, M. & Bolker, B.M. (2017) Modeling zero-inflated count data with glmmTMB.
619 *bioRxiv*, e132753.
- 620 Brunner, J. & Austin, P.C. (2009) Inflation of type I error rate in multiple regression when
621 independent variables are measured with error. *Canadian Journal of Statistics*, **37**, 33–46.
- 622 Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Nave,
623 G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T.Z., Chen, Y.L., Forsell, E.,
624 Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J.,
625 Wagenmakers, E.J. & Wu, H. (2018) Evaluating the replicability of social science
626 experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, **2**,
627 637–644.
- 628 Casella, G. & Berger, R.L. (2002) *Statistical inference*, 2nd edn. Duxbury Press, Pacific Grove,
629 California.
- 630 Cochran, W.G. (1950) The comparison of percentages in matched samples. *Biometrika*, **37**, 256–266.
- 631 Dupont, W.D. & Plummer, W.D. (1998) Power and sample size calculations for studies involving
632 linear regression. *Controlled Clinical Trials*, **19**, 589–601.
- 633 Ebersole, C.R., Atherton, O.E., Belanger, A.L., Skulborstad, H.M., Allen, J.M., Banks, J.B., Baranski,
634 E., Bernstein, M.J., Bonfiglio, D.B.V., Boucher, L., Brown, E.R., Budiman, N.I., Cairo, A.H.,
635 Capaldi, C.A., Chartier, C.R., Chung, J.M., Cicero, D.C., Coleman, J.A., Conway, J.G.,
636 Davis, W.E., Devos, T., Fletcher, M.M., German, K., Grahe, J.E., Hermann, A.D., Hicks,
637 J.A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D.J., Joy-Gaba, J.A., Juzeler, H.,
638 Keres, A., Kinney, D., Kirshenbaum, J., Klein, R.A., Lucas, R.E., Lustgraaf, C.J.N., Martin,
639 D., Menon, M., Metzger, M., Moloney, J.M., Morse, P.J., Prislín, R., Razza, T., Re, D.E.,
640 Rule, N.O., Sacco, T.F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K.,
641 Sternglanz, R.W., Summerville, A., Tskhay, K.O., van Allen, Z., Vaughn, L.A., Walker, R.J.,
642 Weinberg, A., Wilson, J.P., Wirth, J.H., Wortman, J. & Nosek, B.A. (2016) Many labs 3:
643 evaluating participant pool quality across the academic semester via replication. *Journal of*
644 *Experimental Social Psychology*, **67**, 68–82.
- 645 Fordyce, J.A., Gompert, Z., Forister, M.L. & Nice, C.C. (2011) A hierarchical bayesian approach to
646 ecological count data: a flexible tool for ecologists. *Plos One*, **6**, e26785.
- 647 Forstmeier, W., Wagenmakers, E.J. & Parker, T.H. (2017) Detecting and avoiding likely false-
648 positive findings – a practical guide. *Biological Reviews*, **92**, 1941–1968.
- 649 Frank, S.A. (2009) The common patterns of nature. *Journal of Evolutionary Biology*, **22**, 1563–1585.
- 650 Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*, 1
651 edn. Cambridge University Press, New York.
- 652 Ghasemi, A. & Zahediasl, S. (2012) Normality tests for statistical analysis: a guide for non-
653 statisticians. *Int J Endocrinol Metab*, **10**, 486–489.
- 654 Glass, G.V., Peckham, P.D. & Sanders, J.R. (1972) Consequences of failure to meet assumptions
655 underlying the fixed effects analysis of variance and covariance. *Review of educational*
656 *research*, **42**, 237–288.
- 657 Good, P.I. (2005) *Permutation, parametric, and bootstrap tests of hypotheses*, 3 edn. Springer, New
658 York, NY.
- 659 Hack, H.R.B. (1958) An empirical investigation into the distribution of the F-ratio in samples from
660 two non-normal populations. *Biometrika*, **45**, 260–265.
- 661 Harrison, X.A. (2014) Using observation-level random effects to model overdispersion in count data
662 in ecology and evolution. *Peerj*, **2**, e616.
- 663 Harrison, X.A. (2015) A comparison of observation-level random effect and Beta-Binomial models
664 for modelling overdispersion in Binomial data in ecology & evolution. *Peerj*, **3**, e1114.

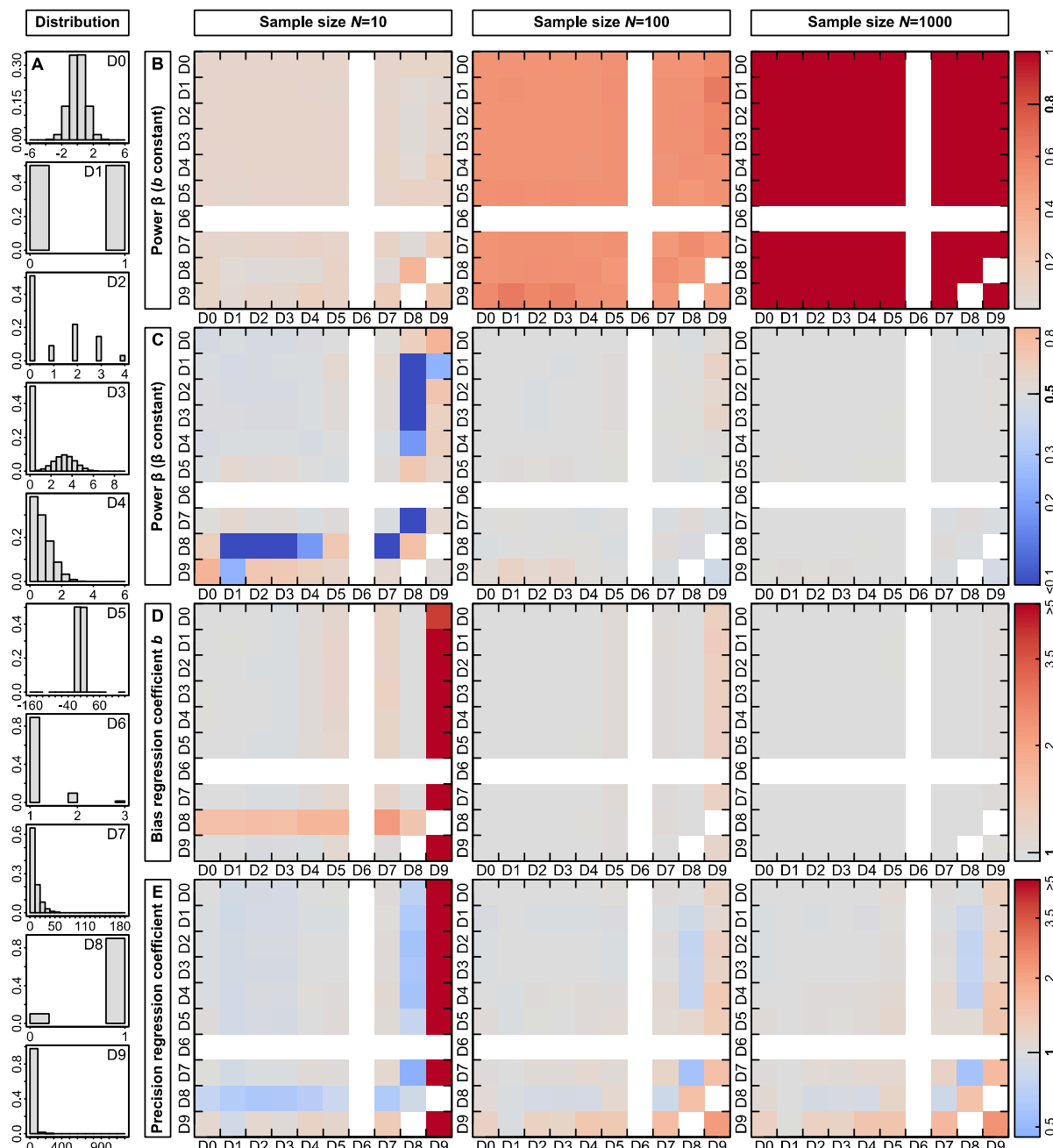
- 665 Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E., Robinson,
666 B.S., Hodgson, D.J. & Inger, R. (2018) A brief introduction to mixed effects modelling and
667 multi-model inference in ecology. *Peerj*, **6**, e4794.
- 668 Hayes, A.F. & Cai, L. (2007) Using heteroskedasticity-consistent standard error estimators in OLS
669 regression: an introduction and software implementation. *Behavior Research Methods*, **39**,
670 709–722.
- 671 Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions.
672 *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*,
673 *Volume 1: Statistics*, pp. 221–233. University of California Press, Berkeley, Calif.
- 674 Ihle, M., Pick, J.L., Winney, I.S., Nakagawa, S. & Burke, T. (2019) Measuring up to reality: null
675 models and analysis simulations to study parental coordination over provisioning offspring.
676 *Frontiers in Ecology and Evolution*, **7**, e142.
- 677 Ives, A.R. (2015) For testing the significance of regression coefficients, go ahead and log-transform
678 count data. *Methods in Ecology and Evolution*, **6**, 828–835.
- 679 Ives, A.R. & Garland, T. (2014) Phylogenetic regression for binary dependent variables. *Modern
680 phylogenetic comparative methods and their application in evolutionary biology* (ed. L.Z.
681 Garamszegi), pp. 231–261. Springer, Berlin, Heidelberg.
- 682 Kass, R.E., Caffo, B.S., Davidian, M., Meng, X.L., Yu, B. & Reid, N. (2016) Ten simple rules for
683 effective statistical practice. *Plos Computational Biology*, **12**, e1004961.
- 684 Knief, U., Schielzeth, H., Backström, N., Hemmrich-Stanisak, G., Wittig, M., Franke, A., Griffith,
685 S.C., Ellegren, H., Kempnaers, B. & Forstmeier, W. (2017) Association mapping of
686 morphological traits in wild and captive zebra finches: reliable within, but not between
687 populations. *Molecular Ecology*, **26**, 1285–1305.
- 688 Komsta, L. & Novomestky, F. (2015) moments: Moments, cumulants, skewness, kurtosis and related
689 tests.
- 690 Lin, L.I. (1989) A concordance correlation-coefficient to evaluate reproducibility. *Biometrics*, **45**,
691 255–268.
- 692 Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002) The importance of the normality assumption in
693 large public health data sets. *Annual Review of Public Health*, **23**, 151–169.
- 694 Mardia, K.V. (1971) The effect of nonnormality on some multivariate tests and robustness to
695 nonnormality in the linear model. *Biometrika*, **58**, 105–121.
- 696 McGuinness, K.A. (2002) Of rowing boats, ocean liners and tests of the ANOVA homogeneity of
697 variance assumption. *Austral Ecology*, **27**, 681–688.
- 698 Miller, R.G. (1986) *Beyond ANOVA: basics of applied statistics*. John Wiley & Sons, Inc.
- 699 O'Hara, R.B. (2009) How to make models add up—a primer on GLMMs. *Annales Zoologici Fennici*,
700 **46**, 124–137.
- 701 O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and
702 Evolution*, **1**, 118–122.
- 703 Önöz, B. & Bayazit, M. (2012) Block bootstrap for Mann-Kendall trend test of serially dependent
704 data. *Hydrological Processes*, **26**, 3552–3560.
- 705 Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*,
706 **349**, aac4716.
- 707 Osborne, J.W. & Overbay, A. (2004) The power of outliers (and why researchers should ALWAYS
708 check for them). *Practical Assessment, Research & Evaluation*, **9**, Available online:
709 <http://PAREonline.net/getvn.asp?v=9&n=6>.
- 710 Osborne, J.W. & Waters, E. (2002) Four assumptions of multiple regression that researchers should
711 always test. *Practical Assessment, Research & Evaluation*, **8**, Available online:
712 <http://PAREonline.net/getvn.asp?v=8&n=2>.
- 713 Pinheiro, J.C. & Bates, D.M. (2000) *Mixed-effects models in S and S-PLUS*. Springer.
- 714 Plaschke, S., Bulla, M., Cruz-López, M., Gómez del Ángel, S. & Küpper, C. (2019) Nest initiation
715 and flooding in response to season and semi-lunar spring tides in a ground-nesting shorebird.
716 *Frontiers in Zoology*, **16**, e15.
- 717 Puth, M.T., Neuhauser, M. & Ruxton, G.D. (2014) Effective use of Pearson's product-moment
718 correlation coefficient. *Animal Behaviour*, **93**, 183–189.
- 719 Qiu, W. (2018) powerMediation: Power/Sample Size Calculation for Mediation Analysis.

- 720 Quinn, G.P. & Keough, M.J. (2002) *Experimental design and data analysis for biologists*. Cambridge
721 University Press, New York.
- 722 Ramsey, F. & Schafer, D.W. (2013) *The statistical sleuth: a course in methods of data analysis*, 3
723 edn. Brooks/Cole.
- 724 Ruscio, J. & Kaczetow, W. (2008) Simulating multivariate nonnormal data using an iterative
725 algorithm. *Multivariate Behavioral Research*, **43**, 355–381.
- 726 Santema, P., Schlicht, E. & Kempenaers, B. (2019) Testing the conditional cooperation model: what
727 can we learn from parents taking turns when feeding offspring? *Frontiers in Ecology and*
728 *Evolution*, **7**, e94.
- 729 Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F.,
730 Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A.,
731 Dalla Rosa, A., Dam, L., Evans, M.H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M.,
732 Glenz, A., Gordon-McKeon, S., Heaton, T.J., Hederos, K., Heene, M., Hofelich Mohr, A.J.,
733 Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei,
734 R., Lindsay, T.A., Liverani, S., Madan, C.R., Molden, D., Molleman, E., Morey, R.D.,
735 Mulder, L.B., Nijstad, B.R., Pope, N.G., Pope, B., Prenoveau, J.M., Rink, F., Robusto, E.,
736 Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F.D., Sherman, M.F., Sommer, S.A.,
737 Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello,
738 M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S. & Nosek, B.A. (2018) Many analysts, one
739 data set: making transparent how variations in analytic choices affect results. *Advances in*
740 *Methods and Practices in Psychological Science*, **1**, 337–356.
- 741 Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-positive psychology: undisclosed
742 flexibility in data collection and analysis allows presenting anything as significant.
743 *Psychological Science*, **22**, 1359–1366.
- 744 Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*. W. H. Freeman, New York.
- 745 Szöcs, E. & Schäfer, R.B. (2015) Ecotoxicology is not normal. *Environmental Science and Pollution*
746 *Research*, **22**, 13990–13999.
- 747 Warton, D.I. & Hui, F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology.
748 *Ecology*, **92**, 3–10.
- 749 Warton, D.I., Lyons, M., Stoklosa, J. & Ives, A.R. (2016) Three points to consider when choosing a
750 LM or GLM test for count data. *Methods in Ecology and Evolution*, **7**, 882–890.
- 751 Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-
752 Newton method. *Biometrika*, **61**, 439–447.
- 753 White, H. (1980) A Heteroskedasticity-consistent covariance matrix estimator and a direct test for
754 heteroskedasticity. *Econometrica*, **48**, 817–838.
- 755 Williams, M.N., Grajales, C.A.G. & Kurkiewicz, D. (2013) Assumptions of multiple regression:
756 correcting two misconceptions. *Practical Assessment, Research & Evaluation*, **18**, Available
757 online: <http://pareonline.net/getvn.asp?v=18&n=11>.
- 758 Young, L.J., Campbell, N.L. & Capuano, G.A. (1999) Analysis of overdispersed count data from
759 single-factor experiments: a comparative study. *Journal of Agricultural Biological and*
760 *Environmental Statistics*, **4**, 258–275.
- 761 Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. (2009) *Mixed effects models and*
762 *extensions in ecology with R*, 1 edn. Springer, New York.
- 763 Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common
764 statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.
- 765 Zuur, A.K., Ieno, E.N. & Smith, G.M. (2007) *Analysing ecological data*. Springer Science + Business
766 Media, LLC.

767



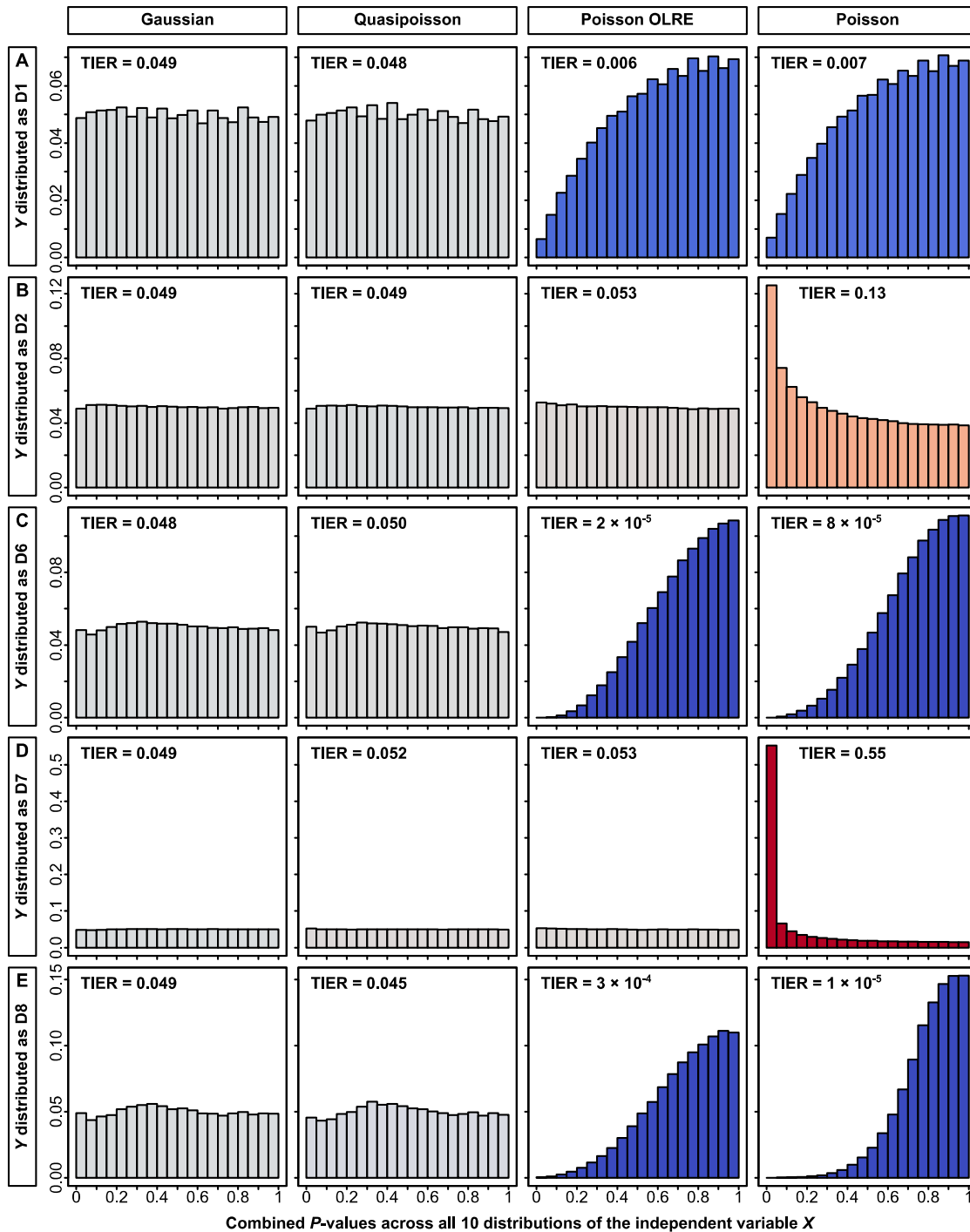
768
 769 **Figure 1** | P -values from Gaussian linear regression models are in most cases unbiased. (A) Overview
 770 of the ten different distributions that we simulated. Distributions D0 is Gaussian and all remaining
 771 distributions are sorted by their tendency to produce strong outliers. Distributions D1, D2, D6, D7 and
 772 D8 are discrete. The numbers D0–D9 refer to the plots in (B–E) where on the Y -axis the distribution
 773 of the dependent variable and on the X -axis of the predictor is indicated. (B) Type I error rate at an α -
 774 level of 0.05 for sample sizes of $N = 10, 100$ and 1000 . Red colours represent increased and blue
 775 conservative type I error rates. (C) Scale shift parameter, (D) bias in P -values at an expected P -value
 776 of 10^{-3} and (E) bias in P -values at an expected P -value of 10^{-4} .



777 **Figure 2** | Power, bias and precision of parameter estimates from Gaussian linear regression models
 778

779 are in most cases unaffected by the distributions of the dependent variable Y or the predictor X . (A)
 780 Overview of the different distributions that we simulated, which were the same as in **Figure 1**. The
 781 numbers D0–D9 refer to the plots in (B–E) where on the Y -axis the distribution of the dependent
 782 variable and on the X -axis of the predictor is indicated. (B) Power at a regression coefficient $b = 0.2$
 783 for sample sizes of $N = 10, 100$ and 1000 . Red colours represent increased power. (C) Power at
 784 regression coefficients $b = 0.59, 0.19$ and 0.06 for sample sizes of $N = 10, 100$ and 1000 , respectively,
 785 where the expected power derived from a normally distributed Y and X is 0.5 . Red colours represent

786 increased and blue colours decreased power. **(D)** Bias and **(E)** precision of the regression coefficient
787 estimates at an expected $b = 0.2$ for sample sizes of $N = 10, 100$ and 1000 .



788
789

Figure 3 | Distribution of observed P -values (when the null hypothesis is true) as a function of

790

different model specifications (columns) and different distributions of the dependent variable Y (rows

791

A to **E**). Each panel was summed up across 10 different distributions of the predictor X (500,000

792

simulations per panel with $N = 100$ data points per simulation). Models were fitted either as glms with

793

a Gaussian error structure that violate the normality assumption (first column), as glms with a

794

Quasipoisson error structure that take overdispersion into account (second column), as glmms with a

795 Poisson error structure and an observation-level random effect (OLRE; Harrison *et al.* 2018) or as
796 glms with a Poisson error structure that violate the assumption of the Poisson distribution. In each
797 panel, TIER indicates the realized type I error rate (across the 10 different predictor distributions),
798 highlighted with a colour scheme as in **Figure 1B** (blue: below the nominal level of 0.05, red: above
799 the nominal level, grey: closely matching the nominal level). The dependent variable Y was
800 distributed as **(A)** distribution D1, **(B)** distribution D2, **(C)** distribution D6, **(D)** distribution D7 or **(E)**
801 distribution D8 (see **Table 1** and **Figure 1A** for details).

802 **Table 1** | Description of the 10 simulated distributions of the independent variable Y and the predictor X .

Name	Sampling distribution	Mean	Variance	Categories	Degree of zero-inflation	Skewness [†]	Kurtosis [†]	Arguments in TrustGauss [§]
D0	Gaussian	0	1	-	0	1.9×10^{-5}	3.00	DistributionY="Gaussian", MeanY.gauss=0, SDY.gauss=1
D1	Binomial	0.5	0.25	-	0	6.5×10^{-6}	1.00	DistributionY="Binomial", zeroLevelY.zero=0.5
D2	Gaussian with categories and zero-inflation [#]	0	1	5	0.5	0.64	2.02	DistributionY="GaussianZeroCategorical", MeanY.gauss=3, SDY.gauss=1, nCategoriesY.cat=5
D3	Gaussian with zero-inflation [#]	0	1	-	0.5	0.45	1.69	DistributionY="GaussianZero", MeanY.gauss=3, SDY.gauss=1, zeroLevelY.zero=0.5
D4	Absolute Gaussian [#]	0	1	-	0	1.00	3.87	DistributionY="AbsoluteGaussian", MeanY.gauss=0, SDY.gauss=1
D5	Student's t	0	2	-	0	0.01	20.71	DistributionY="StudentsT", DFY.student=4
D6	Gamma with categories [#]	10	100	3	0	3.45	15.09	DistributionY="GammaCategorical", nCategoriesY.cat=3, ShapeY.gamma=1, ScaleY.gamma=10
D7	Negative Binomial	10	110	-	0	2.00	9.02	DistributionY="NegativeBinomial", ShapeY.gamma=1, ScaleY.gamma=10
D8	Binomial	0.9	0.09	-	0	-2.67	8.12	DistributionY="Binomial", zeroLevelY.zero=0.90
D9	Gamma	10	1000	-	0	6.32	62.84	DistributionY="Gamma", ShapeY.gamma=0.1, ScaleY.gamma=100

803 [#] Mean and Variance refer to the distributions prior to adding categories, zero-inflation or taking the absolute values.

804 [†] Skewness and kurtosis were estimated from the simulated distributions with 50 million data points using the moments R package (v0.14, Komsta & Novomestky 2015).

805 [§] Here we specified the arguments for the dependent variable Y only. However, the specified values are identical for the independent variable X .

806 **Table 2** | Summary of power, bias and precision of parameter estimates and interpretability from 50,000 simulation runs across the six combinations of the
 807 dependent variable Y and the predictor X . Each combination was either fitted using a Gaussian error structure or the appropriate error structure according to
 808 the distribution of Y (that is either Poisson with a mean of 1 or binomial with a mean of 0.75). The predefined effect was chosen such that a power of around
 809 0.5 was reached (see **Table S2** for details). The column Effect is the mean estimated effect (intercept + slope) after back-transformation.

Distribution of Y	Distribution of X	Error Distribution	Sample size	Power at $\alpha = 0.05$	Power at $\alpha = 0.001$	Mean of slope b	Variance in slope b	CV of slope b	Mean intercept a	Variance in intercept a	CV of intercept a	Effect	Variance in effect
Poisson	Gaussian	Gaussian	100	0.522	0.094	0.200	9.96×10^{-3}	0.498	1.000	9.70×10^{-3}	0.098	1.201	0.023
Poisson	Gaussian	Poisson	100	0.511	0.090	1.228	0.015	0.100	0.976	9.80×10^{-3}	0.101	1.195	0.022
Binomial	Gaussian	Gaussian	100	0.502	0.085	0.085	1.79×10^{-3}	0.500	0.750	1.82×10^{-3}	0.057	0.835	2.84×10^{-3}
Binomial	Gaussian	Binomial	100	0.504	0.091	0.617	3.63×10^{-3}	0.098	0.762	2.03×10^{-3}	0.059	0.834	2.75×10^{-3}
Poisson	Gamma	Gaussian	100	0.588	0.162	0.023	1.28×10^{-4}	0.502	0.776	1.28×10^{-4}	0.176	0.798	0.017
Poisson	Gamma	Poisson	100	0.537	0.095	1.019	7.67×10^{-5}	0.009	0.818	7.67×10^{-5}	0.142	0.833	0.013
Binomial	Gamma	Gaussian	100	0.459	0.029	0.008	1.55×10^{-5}	0.481	0.669	4.12×10^{-3}	0.096	0.677	3.75×10^{-3}
Binomial	Gamma	Binomial	100	0.549	0.113	0.517	1.15×10^{-4}	0.021	0.634	6.87×10^{-3}	0.131	0.650	5.59×10^{-3}
Poisson	Binomial	Gaussian	100	0.673	0.126	0.534	0.039	0.371	0.599	0.025	0.265	1.133	0.014
Poisson	Binomial	Poisson	100	0.699	0.189	1847.624	1.70×10^{11}	223.359	0.599	0.025	0.264	1.132	0.014
Binomial	Binomial	Gaussian	100	0.510	0.127	0.200	0.012	0.551	0.600	9.96×10^{-3}	0.166	0.800	2.15×10^{-3}
Binomial	Binomial	Binomial	100	0.491	0.094	0.717	0.011	0.146	0.600	0.010	0.167	0.800	2.16×10^{-3}

810