Violating the normality assumption may be the lesser of two evils

Ulrich Knief[1,*] & Wolfgang Forstmeier[2]

[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

Key-words: hypothesis testing, linear model, normality, regression
Running header: Consequences of violating the normality assumption
Word count: ca 3200
Figures & Tables: 1 figure
References: 31
Supplementary Material: 4 figures, 2 tables
Data availability: all scripts bundled in the R package “TrustGauss”

Abstract

1. Researchers are often uncertain about the extent to which it may be acceptable to violate the assumption of normality of errors, which underlies the most-frequently used tests for statistical significance (regression, *t*-test, ANOVA, and linear mixed models with Gaussian error).

2. Here we use Monte Carlo simulations to show that such Gaussian models are remarkably robust to even the most dramatic deviations from normality.

3. We find that *P*-values are generally reliable if either the dependent variable *Y* or the predictor *X* are normally distributed and that bias only occurs if both are heavily skewed (resulting in outliers in both *X* and *Y*). In the latter case, judgement of significance at an α-level of 0.05 is still safe unless sample size is very small. Yet, with more stringent significance criteria as is used when conducting numerous tests (e.g. α = 0.0001) there is a greater risk of making erroneous judgements.

4. Generally we conclude that violating the normality assumption appears to be the lesser of two evils, when compared to alternative solutions that are either unable to account for levels of non-independence in the data (most non-parametric tests) or much less robust (e.g. Poisson models which require control of overdispersion and sophisticated resampling). We argue that the latter may pose a more substantial threat to the reliability of research findings when pragmatically acknowledging that, in the majority of publications, statistical expertise is limited.

Introduction

In the biological, medical and social sciences, the validity of research findings is generally assessed via statistical significance tests. Valid significance tests ensure the trustworthiness of scientific results and should reduce the amount of random noise entering the scientific literature. Brunner and Austin (2009) even regard this as the "primary function of statistical hypothesis testing in the discourse of science". A *P*-value of < 0.05 is usually accepted as sufficiently low for rejecting the null hypothesis. However, the validity of parametric significance tests depends on the whether model assumptions are violated.

In a growing body of literature, researchers express their concerns about irreproducible results (Open Science Collaboration 2015; Ebersole *et al.* 2016; Camerer *et al.* 2018; Silberzahn *et al.* 2018) and it has been argued that the inappropriate use of statistics is a leading cause of irreproducible results (Forstmeier, Wagenmakers & Parker 2017). Yet researchers may often be uncertain about which statistical practices can be considered as safe and which are prone to yield overconfident conclusions. Searching the literature, we found relatively little pragmatic advice (Box & Watson 1962; Mardia 1971; Lumley *et al.* 2002; Gelman & Hill 2007; O'Hara 2009; Zuur, Ieno & Elphick 2010) on the question of whether and when it may be safe to violate the assumption of normality of errors, which underlies the most commonly used tests for statistical significance (linear models “lm” and linear mixed models “lmm” with Gaussian error, which includes the often more widely known techniques of regression, *t*-test, and ANOVA). How much deviation is tolerable under which circumstances (in terms of sample size and α-threshold)?

We here use Monte Carlo simulations to explore how violations of the normality assumption affect the probability of drawing false-positive conclusions (the rate of type I errors), because these are the greatest concern in the current reliability crisis (Open Science Collaboration 2015). We aim at deriving simple rules of thumb, which researchers can use to judge whether the violation may be tolerable and whether the *P*-value can be trusted. Furthermore, we provide an R package (“TrustGauss”) that researchers can use to explore the effect of specific distributions on the reliability of *P*-values. Counter to intuition, we find that violations are rarely problematic, and we argue that the commonly recommended solutions to the problem (e.g. using non-parametric tests, generalized linear models) may represent a greater threat to the reliability of conclusions because of their lower flexibility or robustness.

The linear regression model and its assumptions

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

Basic statistics texts introduce (about) five assumptions that need to be met for interpreting all estimates from linear regression models safely (validity, independence, linearity, homoscedasticity and normality; Gelman & Hill 2007). Recall that these criteria are concerned with the dependent variable Y, or — to be more precise — the regression error e. The predictor X is usually not considered. We refrain from revisiting all criteria in detail, but want to specifically focus on the normality assumption here, which is usually tested via inspecting the distribution of the dependent variable or of the residuals (Zuur, Ieno & Elphick 2010). Both visual approaches (probability or QQ-plots) and formal statistical tests (Shapiro-Wilk) are commonly applied. Formal tests for normality have been criticized because they have low power at small sample sizes and almost always yield significant deviations from normality at large sample sizes (Ghasemi & Zahediasl 2012). Thus, researchers are left with their intuition to decide how severely the normality assumption is violated and how robust regression is to such violations.

Normally distributed errors are generally assumed to be the least important (yet probably the most widely known) out of the five regression assumptions (Gelman & Hill 2007). Deviations from normality do not bias regression coefficients (Williams, Grajales & Kurkiewicz 2013) and usually do not impair hypothesis testing (no inflated type I error rate, e.g. Bishara & Hittner 2012; Puth, Neuhauser & Ruxton 2014; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016) even at relatively small sample sizes, and with large sample sizes ≥ 500 the Central Limit Theorem guarantees that the test statistic is on average normally distributed (Lumley *et al.* 2002). Importantly, the robustness of regression methods to deviations from normality of the regression errors e does not only depend on sample size, but also on the distribution of the predictor X (Box & Watson 1962; Mardia 1971).

Simulations to assess effects on *P*-values

To illustrate the consequences of violating the normality assumption, we performed Monte Carlo simulations on five continuous and five discrete datasets that were severely skewed, platy- and leptokurtic or zero-inflated (distributions D0–D9; **Figure 1A** left column), going beyond previous studies that examined less dramatic violations (Bishara & Hittner 2012; Puth, Neuhauser & Ruxton 2014; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016). We explored these 10 distributions across a range of sample sizes (N = 10, 25, 50, 100, 250, 500, 1000). Starting with the normal distribution D0 for reference, we sorted the remaining distributions D1–D9 by increasing tendency to produce strong outliers (calculated as the average distance of the maximum or minimum from the sample mean relative to the standard deviation of the sample for the case of N = 100). We used these data both as our dependent variable Y and as our predictor variable X in linear regression models, yielding 10 × 10 = 100 combinations of Y and X for each sample size (see **Figure S1** for distributions of the independent variable Y, the predictor X, and residuals).

We assessed the significance of the models via an F-test wherever possible and used a likelihood ratio test otherwise. We fitted these models to 50,000 datasets for each combination of the dependent and predictor variable. We did not simulate any effect, which means that both the regression coefficient b and the intercept a were on average zero. This enabled us to use the frequency of all models that yielded a P-value ≤ 0.05 as an estimate of the type I error rate at an α-level of 0.05. The null distribution of P-values is uniform on the interval [0,1] and because all P-values are independent and identically distributed, we constructed confidence intervals using a beta-distribution (cf. Casella & Berger 2002; QQ-plots of expected vs observed P-values are depicted in **Figure S1**). We assessed the deviation of observed from expected -$\log_{10}$(P-value) at an expected value of 3 ($P = 10^{-3}$) and 4 ($P = 10^{-4}$) and by estimating the scale shift parameter $\upsilon = \sigma_{\text{observed}} / \sigma_{\text{expected}}$ (Lin 1989), where σ is the variance in -$\log_{10}$(P-value).

Since some of the predictor variables were binary rather than continuous, our regression models also comprise the situation of classical two-sample t-tests, and we assume that the results would also generalize to the situation of multiple predictor levels (ANOVA), which can be decomposed to multiple binary predictors. To demonstrate that our conclusions from univariate models (involving a single predictor) generalize to the multivariate case (involving several predictors), we fitted the above models with a sample size of $N = 100$ to the same 10 dependent variables with three normally distributed predictors and one additional predictor sampled from the 10 different distributions. We further fitted the above models as mixed-effects models using the lme4 R package (v 1.1-14; Bates *et al.* 2015). For that we simulated $N = 100$ independent samples each of which was sampled twice, such that the single random effect “sample ID” explained roughly 30% of the variation in *Y*. We encourage readers to try their own simulations using our R package “TrustGauss”.

Results

The rate at which linear regression models with Gaussian error structure produced false-positive results (type I errors) was very close to the expected value of 0.05 (**Figure 1B).** When sample size was high ($N = 1000$), type I error rates ranged only between 0.044 and 0.052, across the 100 combinations of distributions of the dependent variable *Y* and the predictor *X*. Hence, despite of even the most dramatic violations of the normality assumption (see e.g. distributions D4 and D9 in **Figure 1A**), there was no increased risk of obtaining false-positive results. At $N = 100$, the range was still remarkably narrow (0.037–0.058), and only for very low sample sizes ($N = 10$) we observed 4 out of 100 combinations which yielded notably elevated type I error rates in the range of 0.086 to 0.11. These four cases all involved combinations of the distributions D4 and D9, which yield extreme outliers. For this low sample size of $N = 10$, there were also cases where type I error rates were clearly too low (down to 0.015, involving distributions D1–D3 where extreme values are rarer than under the normal distribution D0; for details see **Table S1**).

Next we examine the scale shift parameter (**Figure 1C**) which evaluates the match between observed and expected *P*-values across the entire range of *P*-values (not only the fraction at the 5% cut-off). Whenever either the dependent variable *Y* or the predictor *X* was normally distributed, the observed and expected *P*-values corresponded very well (first row and first column in **Figure 1C**).

Accordingly, the *P*-values fell within the 95% confidence bands across their entire range (rightmost column in **Figures S1**). This observation was unaffected by sample size (**Table S2**). However, if both the dependent variable *Y* and the predictor *X* were heavily skewed, consistently inflated *P*-values outside the confidence bands occurred, yet this was almost exclusively limited to the case of $N = 10$ (**Figure 1C)**. For larger sample sizes only the most extreme distribution D9 produced somewhat unreliable *P*-values (**Figure 1C)**. This latter effect of unreliable (mostly anti-conservative) *P*-values was most pronounced when judgements were made at a very strict α-level (**Figure 1D** $\alpha = 0.001$ and **Figure 1E** $\alpha = 0.0001$). At a sample size of $N = 100$, and for $\alpha = 0.001$, observed $-\log_{10}$(*P*-values) were biased maximally 3.36-fold when both *X* and *Y* were sampled from distribution D9. This means 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 $N = 100$, and for $\alpha = 0.0001$, the bias was maximally 4.54-fold (**Figure 1E**). Our multivariate and mixed-model simulations confirmed that these patterns are general and also apply to models with multiple predictor variables (**Figure S3**) and to models with random effects (**Figures S4**).

In summary, *P*-values from such Gaussian models are highly robust to even substantial violation of the normality assumption and can be trusted, except when involving distributions with extreme outliers (distribution D9). For very small sample sizes, judgements should preferably be made at $\alpha = 0.05$ (rather than at more strict thresholds) and should also beware of outliers in both *X* and *Y*.

Drawbacks of alternative solutions

When the assumption of normality of errors is not met, it is often recommended to switch to either non-parametric tests (e.g. Spearman rank correlation, Wilcoxon signed-rank test) or to model a more specific error structure in a generalized linear model “glm” (e.g. binomial, negative binomial, Poisson, zero-inflated Poisson). How risky are these approaches in terms of yielding type I errors?

In contrast to Gaussian models, for instance Poisson models are not at all robust to violations of the distribution assumption. For comparison, we fitted the above univariate models involving the five discrete distributions (D1, D2, D4, D6, D8) with a sample size of $N = 100$ using a Poisson error structure. This yielded heavily biased type I error rates (at $\alpha = 0.05$) in either direction ranging from 0 to as high as 0.56, (**Figures S2**). Such inflations of type I error rates in glms have been reported frequently (Warton & Hui 2011; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016) and this problem threatens the reliability of research whenever such models are implemented with insufficient statistical expertise. First, it is absolutely essential to control for overdispersion in the data, which may be particularly strong when Poisson errors are applied to measurements of areas (e.g. counts of pixels or mm^2), latencies (e.g. counts of seconds), or concentrations (e.g. counts of molecules), besides the more classical abundances (e.g. counts of animals). Failure to account for overdispersion will typically result in very high rates of type I errors (Warton & Hui 2011; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016; Forstmeier, Wagenmakers & Parker 2017). Second, even after accounting for overdispersion, some models may still yield inflated type I error rates, therefore requiring statistical testing via a resampling procedure (Warton & Hui 2011; Ives 2015; Szöcs & Schäfer 2015; Warton *et al.* 2016). While most statistical experts might advocate for such a sophisticated approach to count data, we are concerned about practicability when non-experts have to make decisions about the most adequate resampling procedure. In this field of still developing statistical approaches it seems much easier to get things wrong (and obtain a highly overconfident *P*-value) than to get everything right. Finally, with the inclusion of random effects glmms are much more computationally intensive than lmms and often fail to converge, leading to the recommendation to model all traits as Gaussian (e.g. Ives & Garland 2014).

The biggest downside of non-parametric approaches is that they are less advanced and user-friendly compared to linear (mixed) models (e.g. Akritas & Brunner 2003), such that only simple procedures are widely known and applied. The latter, however, are applicable only to the simplest and idealized scenario of fully independent data points and of only a single explanatory variable with no confounding factors or covariates to be controlled for. Real data sets rarely fulfil that condition, such that simple non-parametric tests often suffer from pseudoreplication and unaccounted confounds. Pseudoreplication, i.e. overestimation of the number of truly independent replicates, results in overconfident estimates and hence is one of the leading causes of false-positive conclusions

(Forstmeier, Wagenmakers & Parker 2017). Gaussian models, in contrast, allow us to easily control for pseudoreplication by specifying the random effects that cause non-independence of data points (mixed-effects models).

Finally, there is much to be gained when researchers follow a standardized way of reporting effect sizes (Lumley *et al.* 2002). For instance, a study that examines the effect of a single treatment on multiple dependent variables (e.g. health parameters) may often switch forth and back between reporting parametric and non-parametric test statistics depending on how strongly the trait of interest deviates from normality, rendering a comparison of effect sizes difficult. Methods of converting effect sizes for discrete traits (e.g. odds ratio from a 2×2 contingency table) into effect sizes for continuous traits (e.g. Pearson correlation coefficient) already work by violating the normality assumption (e.g. fitting a Pearson correlation through the binary data of a 2×2 table; Nakagawa & Cuthill 2007), so why not always report the Gaussian model to begin with, if the primary purpose of the test is to obtain a reliable *P*-value?

Practical advice (for referees)

In order to effectively guard against false-positive claims entering the scientific literature, violations of the normality assumption in linear models are much less of a problem than violations of the independence of data points (pseudoreplication; Schielzeth & Forstmeier 2009; Forstmeier, Wagenmakers & Parker 2017). To avoid the negative consequences of strong deviations from normality that may occur under some conditions (see **Figure 1**) it may be most advisable to apply a rank-based inverse normal (RIN) transformation (aka rankit scores Bliss 1967) to the data, which can approximately normalize most distributional shapes and which effectively minimizes type I errors and maximises statistical power (Bishara & Hittner 2012).

In practice, we recommend the following to referees:

(1) When a test assumes Gaussian errors, request a check for outliers, particularly if very small *P*-values are reported. Consider recommending a RIN-transformation for strong deviations from normality.

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

(3) Requesting a switch to non-parametric statistics is not advised, and requests for switching from lm to glm (or from lmm to glmm) should be accompanied with sufficient advice (e.g. R-code) to ensure a safe implementation.

Conclusion

If we are interested in statistical hypothesis testing, linear regression models with Gaussian error structure are generally robust to violations of the normality assumption. Judging *P*-values at the threshold of $\alpha = 0.05$ is nearly always safe, but if both *Y* and *X* are skewed, we should avoid being overly confident in very small *P*-values and examine whether these result from outliers in both *X* and *Y* (see also Osborne & Overbay 2004). With this caveat in mind, violating the normality assumption is relatively unproblematic. Alternative solutions like Poisson models and non-parametric tests may bear a greater risk of yielding anti-conservative *P*-values when applied by scientists with limited statistical expertise.

Data availability

All functions are bundled in an R package named “TrustGauss”. The R package and all supplementary figures are accessible through the Open Science Framework (osf.io/r5ym4).

Acknowledgements

We thank S. Nakagawa, M. Neuhäuser, and H. Schielzeth for helpful discussions and B. Kempenaers and J.B.W. Wolf for their support.

Author contributions

WF and UK conceived of the study. UK wrote the simulation code. UK and WF prepared the manuscript.

Competing interests

The authors declare no competing financial interests.

References

Akritas, M.G. & Brunner, E. (2003) Nonparametric models for ANOVA and ANCOVA: a review. *Recent Advances and Trends in Nonparametric Statistics***,** 79–91.

Ali, M.M. & Sharma, S.C. (1996) Robustness to nonnormality of regression F-tests. *Journal of Econometrics*, **71,** 175–205.

Bates, D., Mächler, M., Bolker, B.M. & Walker, S.C. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67,** 1–48.

Bishara, A.J. & Hittner, J.B. (2012) Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, **17,** 399–417.

Bliss, C.I. (1967) *Statistics in biology*. McGraw-Hill, New York, NY.

Box, G.E.P. & Watson, G.S. (1962) Robustness to non-normality of regression tests. *Biometrika*, **49,** 93–106.

Brunner, J. & Austin, P.C. (2009) Inflation of type I error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics*, **37,** 33–46.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T.Z., Chen, Y.L., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.J. & Wu, H. (2018) Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, **2,** 637–644.

Casella, G. & Berger, R.L. (2002) *Statistical inference*, 2nd edn. Duxbury Press, Pacific Grove, California.

Ebersole, C.R., Atherton, O.E., Belanger, A.L., Skulborstad, H.M., Allen, J.M., Banks, J.B., Baranski, E., Bernstein, M.J., Bonfiglio, D.B.V., Boucher, L., Brown, E.R., Budiman, N.I., Cairo, A.H., Capaldi, C.A., Chartier, C.R., Chung, J.M., Cicero, D.C., Coleman, J.A., Conway, J.G., Davis, W.E., Devos, T., Fletcher, M.M., German, K., Grahe, J.E., Hermann, A.D., Hicks, J.A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D.J., Joy-Gaba, J.A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R.A., Lucas, R.E., Lustgraaf, C.J.N., Martin, D., Menon, M., Metzger, M., Moloney, J.M., Morse, P.J., Prislin, R., Razza, T., Re, D.E., Rule, N.O., Sacco, T.F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R.W., Summerville, A., Tskhay, K.O., van Allen, Z., Vaughn, L.A., Walker, R.J., Weinberg, A., Wilson, J.P., Wirth, J.H., Wortman, J. & Nosek, B.A. (2016) Many labs 3: evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, **67,** 68–82.

Forstmeier, W., Wagenmakers, E.J. & Parker, T.H. (2017) Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, **92,** 1941–1968.

Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*, 1 edn. Cambridge University Press, New York.

Ghasemi, A. & Zahediasl, S. (2012) Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab*, **10,** 486–489.

Ives, A.R. (2015) For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution*, **6,** 828–835.

Ives, A.R. & Garland, T. (2014) Phylogenetic regression for binary dependent variables. *Modern phylogenetic comparative methods and their application in evolutionary biology* (ed. L.Z. Garamszegi), pp. 231–261. Springer, Berlin, Heidelberg.

Lin, L.I. (1989) A concordance correlation-coefficient to evaluate reproducibility. *Biometrics,* **45,** 255–268.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002) The importance of the normality assumption in large public health data sets. *Annual Review of Public Health,* **23,** 151–169.

Mardia, K.V. (1971) The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika,* **58,** 105–121.

Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews,* **82,** 591–605.

O'Hara, R.B. (2009) How to make models add up—a primer on GLMMs. *Annales Zoologici Fennici,* **46,** 124–137.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science,* **349,** aac4716.

Osborne, J.W. & Overbay, A. (2004) The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research & Evaluation,* **9,** Available online: http://PAREonline.net/getvn.asp?v=9&n=6.

Puth, M.T., Neuhauser, M. & Ruxton, G.D. (2014) Effective use of Pearson's product-moment correlation coefficient. *Animal Behaviour,* **93,** 183–189.

Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology,* **20,** 416–420.

Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A., Dalla Rosa, A., Dam, L., Evans, M.H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T.J., Hederos, K., Heene, M., Hofelich Mohr, A.J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei, R., Lindsay, T.A., Liverani, S., Madan, C.R., Molden, D., Molleman, E., Morey, R.D., Mulder, L.B., Nijstad, B.R., Pope, N.G., Pope, B., Prenoveau, J.M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F.D., Sherman, M.F., Sommer, S.A., Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S. & Nosek, B.A. (2018) Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science,* **1,** 337–356.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*. W. H. Freeman, New York.

Szöcs, E. & Schäfer, R.B. (2015) Ecotoxicology is not normal. *Environmental Science and Pollution Research,* **22,** 13990–13999.

Warton, D.I. & Hui, F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology,* **92,** 3–10.

Warton, D.I., Lyons, M., Stoklosa, J. & Ives, A.R. (2016) Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution,* **7,** 882–890.

Williams, M.N., Grajales, C.A.G. & Kurkiewicz, D. (2013) Assumptions of multiple regression: correcting two misconceptions. *Practical Assessment, Research & Evaluation,* **18,** Available online: http://pareonline.net/getvn.asp?v=18&n=11.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution,* **1,** 3–14.

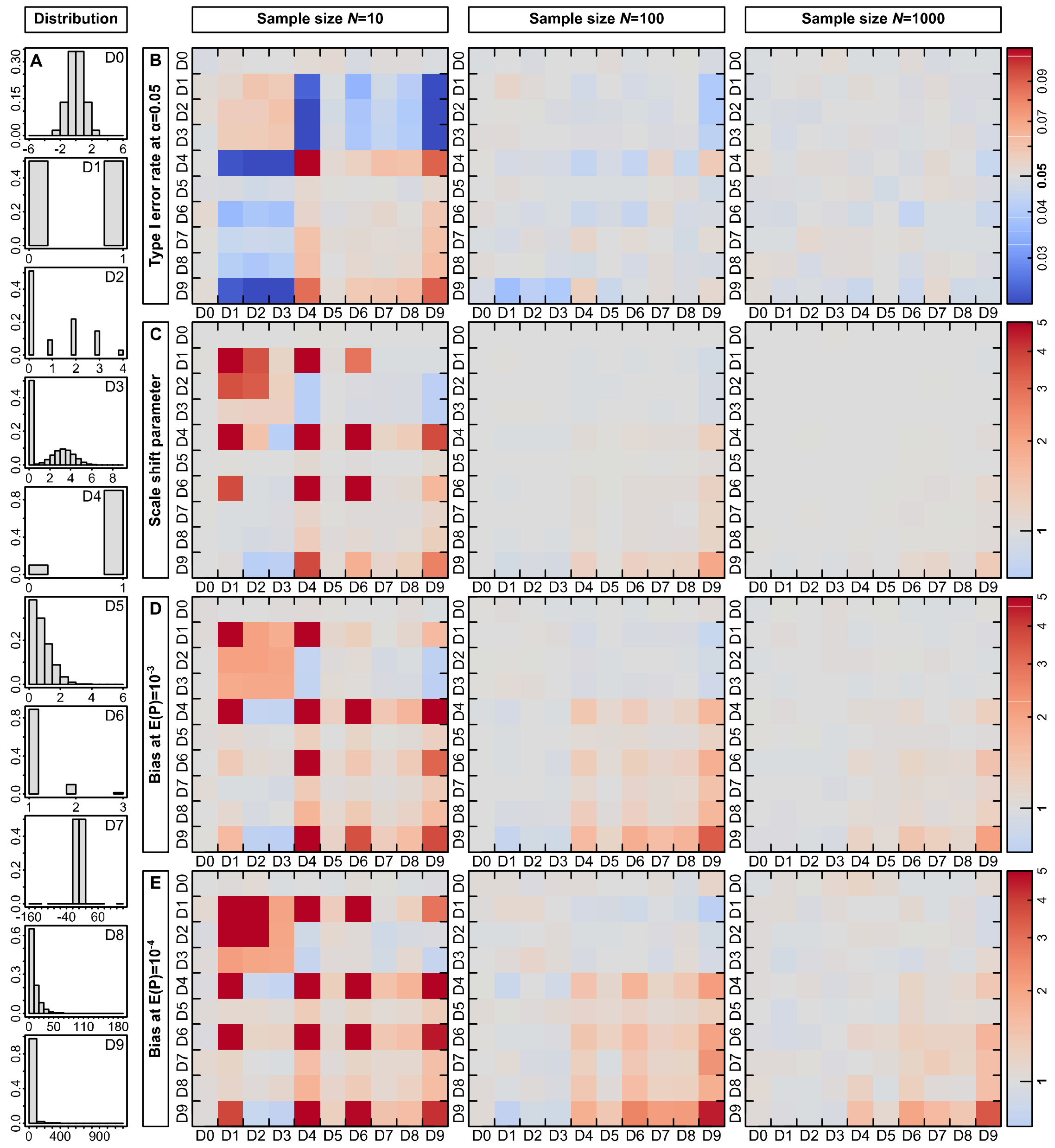

Figure 1 | *P*-values from Gaussian linear regression models are in most cases unbiased. (**A**) Overview of the ten different distributions that we simulated. Distributions D0 is Gaussian and all remaining distributions are sorted by their tendency to produce strong outliers. Distributions D1, D2, D4, D6 and D8 are discrete. The roman numbers refer to the plots in (**B–E**) where on the *Y*-axis the distribution of the dependent variable and on the *X*-axis of the predictor is indicated. (**B**) Type I error rate at an α-level of 0.05 for sample sizes of *N* = 10, 100 and 1000. Red colours represent increased and blue conservative type I error rates. (**C**) Scale shift parameter, (**D**) the bias in *P*-values at an expected *P*-value of 10^{-3} and (**E**) the bias in *P*-values at an expected *P*-value of 10^{-4}.