A few words on the use of statistics by ecologists and evolutionary biologists.

EBD-CSIC Discussion group

"He uses statistics as a drunken man uses lamp-posts... for support rather than illumination." - Andrew Lang (1844-1912)

Statistics is a complex and dynamic field. Ecologists and evolutionary biologists are power users of statistical methods to analyze and interpret data. Hence, a correct understanding and interpretation of statistical tests or models are basic to our endeavor. However, with the increasing pressure to publish and to provide clear messages that are attractive to editors and readers, statistics are susceptible to become a double-edged sword. In fact, every year there are several publications highlighting the perils of focusing almost exclusively on P-values, and the fallacy and undesirable consequences of using statistical significance as a dichotomy to classify true and false hypothesis (see further details in Cohen 1994). The main criticisms to P-values are (i) the use of arbitrary thresholds (e.g. 0.05), (ii) that they depend on too many factors, especially sample size, and (iii) that even experienced researchers have problems correctly interpreting them (Greenland et al 2016).

How to avoid falling into the trap of misusing statistical tests? There is not one single solution, but we here list a set of what we consider good practices to share among EBD-CSIC researchers. This document is intended to be a dynamic reflection of current views in statistical analyses, not a mandatory how-to guideline, and is open for updates from the scientific community at EBD.

1) Despite the ruthless evaluation rules imposed by the system, we should strive to generate honest, enduring scientific articles where we report our complete view of our results, never prioritizing productivity over correctness.

2) Statistics is a tool, not an end. Use it correctly, focusing on your question and how to best answer it.

3) Statistical tests can hardly fix a bad experimental design or poorly collected data. Let's be open to discuss our sampling or experimental designs with colleagues and in the case of predoctoral researchers, capitalize on the supervising committees. Even if experimental design is correct, quite often the sample sizes used are insufficient to answer some of the intended questions. Simulating different experimental designs and sample sizes or doing power analyses before starting data collection can greatly help us to collect the right amounts of data to answer our questions.

4) When conducting experiments, use double blind procedures to prevent biases so the person conducting assays or observations is unaware of the genotype, experimental treatment, population of origin, etc, of the individuals or samples being processed during data acquisition . Use both positive and negative controls as much as possible, interspersed amongst the unknown samples or test subjects.

5) Always plot your data, even before running tests (Zuur et al 2010).

6) Provide the full details of your statistical tests. For frequentist tests, this includes not only a P-value, but the sample size, estimates and associated errors (SE or CI), coefficient of determination (r2), and effect size (typically the difference of means between two groups or the strength of the correlation between two variables). Interpret all of them holistically.

7) Be clear on your goals. Are you doing exploratory analysis, null hypothesis testing, assessing the plausibility of different models (i.e. model selection; Ward et al. 2010) or interested in the model predictive power? All options are fine and require different approaches.

8) Be aware of researcher's degrees of freedom (a.k.a p-hacking): https://en.wikipedia.org/wiki/Researcher_degrees_of_freedom.

9) Understand the statistical test you are performing, especially its underlying assumptions, and beware of the default parameters in R. Running a script doesn't mean that you fully understand how it works and how results should be interpreted.

10) Always check model assumptions (e.g. normality, homogeneity of variance and linearity for ANOVA type analysis). Visual checks are often fine. Plot models and data to better interpret them. clearly explain all steps done in methods.

11) P-values are used in null hypothesis testing. This implies you should have some (clear if possible!) hypotheses prior to running the test. So the ideal sequence is to first formulate the hypothesis and second to test it, not in getting a myriad p-values first and then try to explain what may be going on.

12) Be aware of P-values. P-values assess how consistent our observations are with the null hypothesis being true. However, the p-value cannot differentiate if:
    a) The null hypothesis is false
    b) The null hypothesis is true, but our data does not represent well the population (e.g. bias or low sample size).

    In other words, a 'significant' p-value ($p<0.05$) does not mean that the null hypothesis is necessarily false, and a 'non-significant' p-value ($p>0.05$) does not mean that the null hypothesis is true. To unravel such question, we need to repeat the study or experiment multiple times. For example, when statistical power is limited, a repetition of the same

study will likely provide substantial different p-value, regardless of whether the p-value returned from a statistical test is low or high. A relatively low p-value (e.g. 0.01) may give us a false sense of security when sample size is not adequate (Hasley et al. 2015). For instance, Cumming (2008) showed that for a study obtaining a p-value of 0.03, there is a 90% chance that a replicate study would return a p-value between 0 - 0.6. Overall, It is thus urgent to consider the replicability of the p-value and replicate those studies with low statistical power (e.g. Camerer et al. 2016 or Open Science Collaboration). See further details about the p-value and the error rates in Sellke et al. 2001 and Greenland et al. 2016. However, P-values are not to be entirely discarded. They can be especially useful in reporting results from experimental data, but contextualized with the other statistical results (see 5).

13) Think about the statistical power of your analysis. The power is a function of your sample size, the effect size and the variance expected. Calculating the power of an analysis can be extremely difficult (especially without preliminary independent data), but do not trust analysis with suspected low power. There is no specific amount of data to be considered a small or big sample size. The sample size would be relatively small or big to answer our questions depending on the complexity of the model (i.e. amount of parameters) and the effect sizes and spread of the data. For that reason the same amount of data can be enough to answer some questions but not others.

14) Beware of zeros. In ecology we tend to have too many zeros in our data, for example, with count data where a species has not been observed in certain locations during a survey. This could lead to biased parameter estimates and errors. Consider using zero-inflated or hurdle models.

15) Predictive models should not only be assessed statistically (R2, ROC curve, calibration), but also by using independent training and test data. Beware of overfitting (i.e. models with high R2 for your dataset, but non-replicable).

16) When models are developed for decision-making, we should not only consider the statistical performance of the models, but also the associated benefits and costs associated with the predictions (Vickers & Elkin 2006). For example, medicine and ecology may consider useful models with very different error rates. In fact, because of the usefulness of the models partially depend on the costs and benefits associated with accurate predictions, the same model can be useful for some users but not others.

17) Bayesian analyses do not rely on p-values, but you can easily fall into the same dichotomy by looking at CI and its overlap with zero. Once again, see (5).

18) Biological significance (effect size) is more relevant than statistical significance

Recommended references:

Cohen J. 1994. The Earth is Round p<.05. American Psychologist 49:997-1003.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62-71.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods*, *12*(3), 179.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300.

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of peace research*, *47*(4), 363-375.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565-574.

Krzywinski, M., & Altman, N. (2013). Points of significance: error bars.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

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(1), 3-14.

Steel, E. A., M. C. Kennedy, P. G. Cunningham, and J. S. Stanovick. 2013. Applied statistics in ecology: common pitfalls and simple solutions. Ecosphere 4(9):115. http://dx.doi.org/10.1890/ES13-00160.1

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science, 349(6251), p.aac4716.

Halsey, L.G., 2019. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?. Biology letters, 15(5), p.20190174.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. and White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution, 24(3), pp.127-135.

Nakagawa, S. and Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in ecology and evolution, 4(2), pp.133-142.

Ho, J., Tumkaya, T., Aryal, S., Choi, H. and Claridge-Chang, A., 2019. Moving beyond P values: data analysis with estimation graphics. Nature Methods, 16(7), pp.565-566.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), pp.337-350.

Makin, T.R. and de Xivry, J.J.O., 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife, 8.