Annex IX: Statistical significance and effect size

Different GzLMM appear to provide consistent inference and good sensitivity in detecting the significance of effects, but how should 'significance' be properly understood when making inference?

P-values indicate the probability of the observed data (or, more exactly, of data at least as extreme, i.e. divergent, as those actually observed) assuming the null hypothesis is true (together with other underlying assumptions). A *P*-value is a continuous, or graded, evidence for the falsification (i.e., disproval) of the null hypothesis, in the sense that if the null hypothesis were true, we would expect to see sample data like those observed, or even more extreme, with probability *P* (Carver, 1978; Amrhein et al., 2017). Thus, if *P* is small, we would expect to see such a divergence among means, or even greater, only on rare occasions. If we observed data with a low *P*-value, therefore, we are alerted that, assuming the data are right, the null hypothesis is suspect.

Looking at science as a falsification endeavour, we can say that smaller *P*-values raise more doubt against the null hypothesis than larger *P*-values (Amrhein et al., 2017). Hence, the *P*-value should be viewed as a descriptive statistic about the probability that the sample data are observed in a population characterized by the null hypothesis, not like a formal evidence for a dichotomous judgment (Amrhein et al., 2017). As this probability is based on the expected distribution of sampling fluctuations when the null hypothesis is true, *P*-values are expected to fluctuate as well when independent tests are conducted on different samples drawn from a common population, for example in replicate experiments (Carver, 1978; Halsey et al., 2015).

Reducing *P*-values to 'significant' and 'non-significant' is, thus, a frequent cause for which studies seem irreproducible (Amrhein et al., 2017). Significance thresholds (that is, conventional *P*-values below which significance is claimed) were devised to protect researchers from making claims about the relevance of effects that instead are just noise, but dichotomous threshold thinking cannot and must not replace informed judgment (Amrhein et al., 2017). Most importantly, non-significance cannot be interpreted as demonstrating that "there is no effect" (Amrhein et al., 2017); rather, failure to reject the null hypothesis implies only that the studied effect, if any, is not large enough to be detected as significant with the sample size employed in the described experiment.

Albeit with all these caveats, it is here still suggested that actual *P*-values are reported, and significance thresholds are adopted to pragmatically summarize the chief findings, for reasons of general understanding and clarity. In this paper, 'significant' is used with reference to the conventional significance threshold $P \leq 0.05$, which is deemed to be a sensible choice in the context of germination, like in many biological fields, to call for finding an effect that is large enough and consistent enough, under the experiment conditions, to deserve a public notice. It identifies, in other terms, results that accrue a moderate strength of evidence to the disproval of the null hypothesis, providing, however, the null hypothesis has not a very high prior probability of being true (Goodman, 2001; Krzywinski and Altman, 2013). No definitive pronouncement can, anyway, be made only on the basis of the significance or non-significance of an effect in a single experiment: as what is tested is the significance of the divergence of the observed data from the given null hypothesis, only consistent divergences across several independent experiments can prove the implausibility of the null hypothesis.

It ought to be noted that, even though not proving the null hypothesis, publishing non-significant results for well-devised experiments has statistical value, at least if the scope of the study is conceptually meaningful. If, in fact, many experimental trials are done and only those that produce significant results are published, statistical significance could be finally found, even for effects that are not real, particularly if small sample sizes are used (Amrhein et al., 2017). Moreover, if replicate experiments are performed based on small samples, substantial variation in the *P*-value will be found, thus that replicability of *P*-values is poor when sample size is small (Halsey et al., 2015). On the other hand, statistical significance could be (correctly) found for most plausible effects with a sample size that is large enough (Amrhein et al., 2017). However, with very

large sample sizes, it is possible to obtain statistically significant differences that are of trivial interest in reality (Sileshi, 2012). In this sense, quite often a significance test is actually testing whether the sample size is large enough (and/or the experimental setup is refined enough) to detect the effect, rather than whether the effect exists or not (Amrhein et al., 2017). This still makes a lot of sense if we aim to determine if an effect is present that can be detected as 'significant' within a sensible experimental setup, including a reasonable sample size (Krzywinski and Altman, 2013).

Whether an effect appears to be significant or not for germination in a given experiment is the chief question in many studies. Although the importance of the effect size (i.e., the magnitude of the variation in the response among group means) has been emphasised (Carver, 1978; Sileshi, 2012; Stroup, 2015), estimating the exact size of the effect is usually less of concern in the context of germination studies than it is in other fields. Most studies, indeed, focus on hypothesis testing, rather than on estimating the effect size (Sileshi, 2012). This is owing to some trivial reasons.

First, if the effect size is thought of as being on an interval scale (that is, a difference in the response between the levels of a factor), a given size of the effect on the probit scale, where the effect is sized and tested (and where it is supposed to act quantitatively, in the linear model), translates into a broadly uneven size on the percentage scale, where it is of interest. This means that, in germination studies, the practical relevance of the effect depends on the experimental conditions. For example, if a dormancy breaking treatment has an effect size of 0.4 probits, this will correspond to an effect of near 15.5% if the treatment is done on seeds that are already 50% germinating, whereas it will have an incremental effect of only 2.4% if the seeds are instead already 96% germinating (and all viable).

Second, very often the size of the response to a treatment is dependent onto so many factors, intrinsic and extrinsic, that trying to make some general inference about its magnitude outside the tested conditions might be questionable. Consider a fungicide: in an experiment representing quite typical conditions, testing seed-dressing for three seed batches, each with a similar incidence of a given pathogen, shows a consistent improvement of germination from 75% to 95%, with a highly significant effect. Can we make the broad claim that this treatment generally improves germination by 20%? No way: other genotypes, infected with other fungi, at a diverse degree of contamination, under different conditions of aeration, watering and substratum/soil, with a diverse ageing of the seeds, will produce very different effect sizes, not only because the differing conditions affect the efficacy of the treatment, but also because the different conditions could also change the starting level of germination, and this would modify the effect size on the percentage scale, assuming it is constant on the linked scale. However, if the fungicide works well in a number of carefully performed experimental tests, with effects generally significant, it can be safely claimed to have a protective effect on germination (even though they might well vary in size among experiments), maybe better or worse than another chemical, this can be tested as well. How much, however, is a tricky question to answer, and thus providing an answer is a task usually avoided. At least, it is not possible to claim that the observed effect size represents a general inference, in percentage terms. Furthermore, estimates of effect size are highly variable among replicate experiments when small samples are utilized (Halsey et al., 2015).

Greater importance is normally assigned, in germination studies, to the experimental design and setup, and to the tested conditions: if they are sensible, conceptually solid, and are based on a reasonable sample size (Krzywinski and Altman, 2013), it is tacitly assumed that the results can be useful to determine whether the studied effect is 'significant', that is, it is large enough, and consistent enough (across replicates), to have a low probability of being due to chance alone, so that we might expect it will likely be observed under similar conditions. Anyway, 'significant', *per se*, is neither a proof of replicability nor of reliability (Carver, 1978), particularly if small samples are used (Halsey et al., 2015): we just have high hopes for this to happen. Reliable conclusions on the general value of a finding and its replicability can only be drawn once evidence has been accrued from several independent studies (Amrhein et al., 2017). Anyway, using power analysis to preliminary evaluate the probability that an experimental design will find the minimum treatment effect

considered scientifically relevant to be statistically significant (Gbur et al., 2012) is always a laudable approach.

Notwithstanding the above mentioned practical considerations, there are good reasons to ponder the size of the effect too, albeit refraining from making generalizations about its inferred quantification. Just because it is a percentage, or proportion, the germination response to a given effect has a size that is modelled as constant (i.e., linear) on the linked scale, and therefore, on the data scale, it is widest when the effect is determined as a change over the middle (i.e., 50 %) of the percentile range, whereas the effect size declines towards 0% and 100%. It is important to realize that even the precision of inference declines toward the boundaries of the percentile scale even though precision of measurement does not. In fact, since measurements are taken on the data scale (and therefore errors take place on the data scale too), an equitable, or roughly equitable, measurement, or experimental, error on the data scale greatly expands on the linked scale, and statistical significance is assessed on the linked scale.

Consider, for example, an experimental error as small as ±1% throughout the percentile range: though it is practically negligible even on the linked scale for effects measured over the middle of the percentile range (i.e., over 50%), it expands tremendously on the linked scale if the effect is measured toward the boundaries of the percentile range. In fact, a ±1% error around 50% signifies the estimate of the mean ranges from 49% to 51%. On the probit scale, this interval corresponds to about 0.05 probits. However, the same ±1% error around a mean of 2% signifies that the estimate of that mean ranges from 1% to 3% (roughly, as, at the boundaries, the error is restricted to be within the percentile range), which, on the probit scale, corresponds to an interval of 0.45 probits. Thus, the same experimental error is nine times higher for the more extreme percent mean than for the middle range mean, once reported on the linked scale. As the effect size is modelled as constant on the linked scale, it follows that estimating the effect size from data close to the boundaries carries much more uncertainty than if it is estimated over the middle of the percentile range; this uncertainty reflects on any inference based on such data. Hence, researchers need to be wary of greater uncertainty of significance for small effects close to the percentile boundaries. Indeed, an effect size that is significant, on the linked scale, can be very small on the data scale if it is estimated near the boundaries of the percentile range, and, most importantly, its statistical significance could be more aleatory than if obtained as a large percentage difference in the middle of the percentile range.

Since conclusions based on significance alone might be faulty in borderline conditions, judgments also based on the observed effect size (on the original scale) will increase inferential reproducibility (Amrhein et al., 2017). It is therefore sensible that, even if statistically significant, an effect is also evaluated considering what effect size is deemed relevant, on the observation scale, in a given context (Stroup, 2015; Halsey et al., 2015). This also enforces to keep a distinction between statistical significance and relevance: a significant effect might, ultimately, be of neither theoretical nor practical interest (Sileshi, 2012). It is therefore recommended that, in germination studies, the tested conditions are such to ensure that the effect size, if the effect is real, can be large enough to be judged as relevant. In this respect, it is here suggested to consider as relevant, for germination studies, an effect size that is larger than 15 %.

Besides, researchers are encouraged to consider the relevance of a given effect in relation to the accuracy of all the assumptions implied in their experiment. In this regard, I am particularly concerned with studies based on mutants, wherein small, albeit statistically significant, differences in the germination response between the mutant and the wild type are claimed to demonstrate a key role of the mutated gene for germination. I wonder if any mutation that causes a difference smaller than 40% could merit to be defined as ‘key’. Furthermore, the mutation, which, of course, must absolutely be in the same genetic background as the control (unless multiple genotypes are contrasted on the two sides), could have an indirect effect rather than a direct, real one. For example, if the studied mutation (which should be repeatedly backcrossed into the wild-type background, to minimize the risk of multiple mutations and cryptic genetic variation, and to maximize the probability that comparisons are made between isogenic, or pseudo-isogenic, wild-type strains;

Chandler et al., 2013) affects the ripening time of the dispersal unit, but all the genotypes are harvested on the same date, a different degree of dormancy can be attained because either the wild type or the mutant underwent some dry after-ripening.

Analogously, diversity of provenience, processing, storage, threshing and sorting can all cause significant effects on the germination response that can be confounded with the genetic effect, especially if the effect size is small. This is a form of pseudoreplication (Sileshi, 2012). Fluctuations of environmental variables can indeed have an effect *per se*. Suitable controls should therefore always be included. To find out the best model system to conduct a study also involves choosing genetic materials and experimental conditions that maximize the effect size with respect to possible confounding effects. Statistical analysis is never as important as ensuring that the real informational content of the observations is coherent with the biological hypothesis under test. Putting it more informally, if the data are not valid for our purpose, no statistical analysis can save us from reaching an unsupported inference.

It ought to be considered that when germination data are analysed with a theoretically-sound model like hydrotime (Gianinetti and Cohn, 2007), data close to the boundaries, in addition to those at the boundaries, are excluded just because of the stronger imprecision of probit values (which are used for modelling) obtained by these data, and this is indeed associated with the precision of observations on the data scale. Whatever significant effect can be detected below 1% or above 99%, it is most probably to be ignored.

References

Amrhein V., Korner-Nievergelt F. and Roth T. (2017). The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* 5: e3544.

Carver R.P. (1978). The case against statistical significance testing. *Harvard Educational Review* 48: 378-399.

Chandler C.H., Chari S. and Dworkin I. (2013). Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends in Genetics* 29: 358-366.

Gbur E.E., Stroup W.W., McCarter K.S., Durham S., Young L.J., Christman M., West M. and Kramer M. (2012). *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. American Society of Agronomy: Madison, WI, USA.

Gianinetti A. and Cohn M.A. (2007). Seed dormancy in red rice. XII: Population-based analysis of dry-afterripening with a hydrotime model. *Seed Science Research* 17: 253-271.

Goodman S.N. (2001). Of *P*-values and Bayes: a modest proposal. *Epidemiology* 12: 295-297.

Halsey L.G., Curran-Everett D., Vowler S.L. and Drummond G.B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods* 12: 179-185.

Krzywinski M. and Altman N. (2013). Points of significance: Power and sample size. *Nature Methods* 10: 1139-1140.

Sileshi G.W. (2012). A critique of current trends in the statistical analysis of seed germination and viability data. *Seed Science Research* 22: 145-159.

Stroup W.W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal* 107: 811-827.