Lomonosov Moscow State University
Faculty of Biology

Mathematics and Reality: the Confrontation of Rigor and Complexity

The Articles
Memories of A.T. Teriokhin

Elena Budilova Ed.

Editions Soliton
Moscow
2012

629 pp

Birth of ideas and on the remarkable person and research scientist Anatoly Terekhin: mathematics, reality and the confrontation of rigor and complexity for combining probabilities

Thierry De Meeûs and Jean-François Guégan

Foreword

Our work with Professor Anatoli Terekhin, even if one of us (JFG) got some collaborations with him earlier at the end of the 90's when they met during an international conference held in France, really began with our GEMI research group while he was staying in Montpellier during years 1999–2000 and 2002–2003. The research project exposed hereafter took birth during friendly discussions we had. Anatoli was a very friendly colleague and extremely patient with us while explaining difficult mathematical issues or taking into account our objections. The discussions grew to such a point that it became obvious some work to be published should be undertaken. This was achieved in two steps, the later one having been finalized only several weeks before Anatoli's death. The first step was mainly assumed by Anatoli himself where he explored the problem with continuous data [Teriokhin et al., 2007]. The second step had to do with proportional data, that concern more population geneticists, and then the release of our Software *MultiTest* [De Meeûs et al., 2009].

The problem

Combining probabilities of a series of tests obtained for the same null and alternative hypotheses (H_0 and H_1) is a very old and difficult issue that is obscured by the diversity of situations involved, the multiplicity of terminology, the complexity of the problem and hidden H_1' hypotheses. When combining probabilities, several H_1' (that we will define below) can arise, which are different from H_1 of each individual test. Many different researchers with very different approaches are lead to sometimes divergent opinions simply because they do not speak about exactly the same matter. So in this short chapter we will quickly recall the different issues concerned with combining test results, how as biologists (TDM is a population geneticist and JFG is a community ecologist) we dealt with this matter when collaborating with Anatoli on that subject, and what new reality emerged from these considerations we had through the inspiring collaboration we had with him.

Background

It may happen that researchers have to take into account the results obtained from different statistical tests of the same null hypothesis. It is then desirable to combine all tests into a single one in order to make the most accurate decision. This is typically the case when one wants to combine the results from different published articles and obtain a global P-value over all the tests for global decision making or, in population genetics studies, when the statistical results from different kinds of samples must be combined. This is an old story, and the international scientific literature is full of this kind of statistical tests for combining independent probabilities. For instance, it may be desirable to test for the effect of smoking during pregnancy on offspring body size at birth in different environments where size at birth is not expected to be the same, or to test for genetic differentiation between males and females from different independent samples, or to test for genetic differentiation between infected and non-infected host individuals from different populations or between parasites collected from different host species sampled in sympatry in different locations. Let p_1, p_2, ...p_k be the k p-values obtained. The more the number of such tests to be combined is rising the more often a significant P-value has a chance to arise even under H_0. From the opposite perspective, if in the k tests series, none of the individual P-values is above 0.5 but none of it is equal or inferior to 0.05, such a distribution is not expected under H_0, even if no individual test is significant at $\alpha = 0.05$. Under H_0 the k tests are expected to follow a uniform distribution with mean 0.5 and limits [0, 1]. For instance the (completely artificial) series 0.499, 0.499, 0.499, 0.499, 0.499, 0.499 of independent P-values would output a global rejection of H_0 with P-value ≤ 0.016 (computed with *MultiTest* V1.2 available at http://gemi.mpl.ird.fr/SiteSGASS/SiteTDM/Programs; see also http://www.biomedcentral.com/1471-2105/10/443). This is far from intuitive for everybody so it is worthy of note here. From there different situations may arise depending on the independence or not of the different tests in the series, and depending on how H_1' is defined by the analyser or whether or not a global procedure actually exists.

The different available procedures before our work with Anatoli

The oldest method, but apparently not the most often used to our knowledge, was first introduced by Wilkinson [Wilkinson, 1951] and first applied (still to our knowledge) to population genetics data by Prugnolle and his collaborators [Prugnolle et al., 2002]. At a given type I error rate α of say 0.05, if k tests are undertaken under H_0, it is expected that there are about 5% of P-values that should be equal or inferior to 0.05 (by definition). Then an exact binomial test with 0.05 expectation, $k_{0.05}$ success, the number of observed P-values not greater than 0.05 in k trials, should provide the exact probability that a number as great or greater of significant P-values can be observed under the null hypothesis (hence the P-value for the k tests series).

A second test is Fisher's procedure [Fisher, 1970; Manly, 1985], which is simply obtained by a Chi-square test with $2\times k$ degrees of freedom on the quantity:

$$\chi^2 = -2\sum_1^k \ln(p_i) \tag{1}$$

Fisher's method is very popular, in particular in population genetics, for combining independent tests and is the preferred procedure in the most popular *Genepop* software [Raymond and Rousset, 1995; Rousset, 2008]. Fischer's method has also occasionally been used by community ecologists (Hugueny and Guégan, 1997).

Bonferroni and its sequential derivatives [Holm, 1979; Rice, 1989; Benjamini and Hochberg, 2000] was initially obtained by dividing the smallest P-value by k and the second smallest by $k - 1$ and so on. Multiplying the smallest of the k P-values by k, the second by $k - 1$ and so on is equivalent. In that case, and for convenience, the P-value is set to 1 if this product gives a value above this limit. Bonferroni correction is also widely used in population genetics analyses and is for instance routinely proposed for multiple paired tests in *Fstat* software [Goudet, 2001] updated from [Goudet, 1995].

The SGM procedure was proposed by Goudet [1999]. It uses the geometric mean of *P*-values as a statistic and a randomization procedure to test for symmetry around 0.5 (hence the acronym SGM, Symmetry around the Geometric Mean). It was mostly designed for meta-analyses of published data. It indeed gives much more weight to high *P*-values (e.g. above 0.9), which are indeed expected to be rare in such literature due to publication bias [De Meeûs et al., 2009].

In 2005 Whitlock proposed Stouffer's Z-transformed test [Whitlock, 2005]. Each *P*-value p_i is transformed into its standard normal deviate Z_i, which, for instance, can be obtained by the normal inverse function of Excel™.

Z_i is used for the computation of the statistic Z_s [Whitlock, 2005]:

$$Z_s = \frac{\sum_i^k Z_i}{\sqrt{k}} \tag{2}$$

Z_s is then compared to the normal standard distribution (*e.g.* NORMSDIST(Z_s;0;1) in Excel). To our knowledge this procedure has hardly ever been used.

The work we undertook with Anatoli was a generalisation of the Wilkinson's binomial simple principle [Teriokhin et al., 2007].

The generalized binomial or Terekhin's test

The general principle of the generalized binomial test is that under H_0 a uniform distribution of *P*-values, centred on 0.5 and limited by 0 and 1, is expected. In other words, any of *P*-values between 0 and 1 has an equal chance to appear in the series under H_0. Thus, even in the absence of any significant *P*-value in the series (say at $\alpha = 0.05$), if the distribution is biased to values below 0.5, this might reflect a significant signal across the whole series of *P*-values. The generalized binomial test looks after, at a given level of significance α, the probability to obtain as many individual p_i's, inferior or equal to a chosen threshold α' in the series. Any *a priori* chosen threshold *P*-value < 0.5 can theoretically work but Anatoli's simulations suggested that $P_{k/2}$, where $P_{k/2}$ is the $(k/2)^{\text{th}}$ *P*-value of the series ranked in increasing order, provides the best results in most situations.

These different procedures briefly exposed below are not equivalent, not only on the results provided out of the same series of *P*-values, but also in terms of what H_1' really is. We call here H_1' the alternative hypothesis over the k tests series, which is not necessarily the same as H_1 of each individual test. Consequently, each procedure does not apply to all situations that can be met. This is not trivial as illustrated by the difficulties we had to make ourselves clear to the different referees for the two subsequent articles we published [Teriokhin et al., 2007; De Meeûs et al., 2009], and which was for our benefit as it forced us to make it clear for ourselves as well.

The k tests are independent

In that case several H_1' are possible.

The first possible H_1' is H_{1_1}': what tests are significant at the chosen level, taking into account the inflated risk of falsely rejecting H_0? Here the only available procedures are those that lower the level of significance to an «acceptable» value like the sequential Bonferroni. Nevertheless, users should be aware that these procedures all are extremely conservative. Hence, users may be encouraged to prudence while accepting H_0.

The second possible question is H_{1_2}': is there at least one significant test in the series? Though this can be handled by Bonferroni-like procedures, Fisher's procedure is exactly testing for that and is much more powerful than Bonferroni in that situation.

The third possibility arises when H_1' is H_{1_3}': is the k-test series significant as a whole, this is where Stouffer's Z and the generalized binomial may apply. In that case, if the series is very short (two to three tests only) it is wiser using Stouffer's Z. Otherwise both statistical procedures are equivalent in power though the generalised binomial represents a more direct assessment of the significance of the series and has our (not totally fair) preference. Nevertheless, it is extremely important to mention that in this third situation, if an exact global test running directly from the data exists, this global test must be preferred [De Meeûs et al., 2009]. Another very important advantage of the generalized binomial is that it can be used even if the exact values of P-values are not known with certainty (which is often the case for published data). No other procedure shares this property.

The k tests are not independent

This is typically the case of *post-hoc* tests for paired data, like after an ANOVA-like test that outputs a significant result, one wants to know which treatments are different from the others. This is also typically met in population genetics for linkage disequilibrium (LD) tests between paired loci or differentiation tests between pairs of subsamples. In such situations a supplementary problem arises because, when H_0 is false, the different P-values are correlated, even if the signal is small. For instance, if we test LD between pairs for six loci (*e.g.* L1, L2, L3, L4, L5 and L6) there will be 15 possible tests. If L1 and L2 are significantly linked, and if L2 is significantly linked to L6, then L1 will have an increased chance of being significantly linked to L6 as well. For this reason, Fisher's and Stouffer's procedures cannot be used here. For H_{1_1}' and H_{1_2}' only Bonferroni and its sequential extensions can be used. For H_{1_3}', the classical binomial can be used (much more powerful than Bonferroni), *i.e.* compute the exact binomial unilateral probability at level α with k trials and k_α success (number of P-values $\leq \alpha$).

Conclusion

All these advances and subtleties were unknown to us and probably to much of the community of population biologists (at least for those that were not well trained in biostatistics and biomathematics, which is a large part of them) before we undertook these works under the leadership of Anatoli.

Our works with Anatoli have changed our vision on that matter, and it will probably contribute to change the habit of other population biologists in a near future. Perseverance, the quality of continuing with something even though it is difficult, which for sure was one of the intellectual quality of Anatoli, and capacity to influence your own field of research and to disseminate through other fields of expertises definitely are the hallmarks of great personalities to whom Anatoli belonged to. The fact that he left us with the charge to promote (t) his work – he worked with us, and co-authored the writing of the 2009' paper only several weeks before dying – is a legacy we are extremely proud to humbly take care of. But this is when we will need his kind advices to improve ourselves that we will really miss him.

Acknowledgements

Thierry De Meeûs and Jean-François Guégan are financed by the CNRS, IRD and the French School of Public Health. We here thank the CNRS (France and Russia) for providing two subsequent "red-positions fellowships" to Anatoli as a senior research scientist in our research team in Montpellier, a first 7-month fellowship from September 1999 to April 2000 and a second 12-month one from September 2002 to August 2003.

References

1. *Benjamini Y., Hochberg Y.* On the adaptive control of the false discovery fate in multiple testing with independent statistics // Journal of Educational and Behavioral Statistics. 2000. V. 25. P. 60–83.
2. *De Meeûs T., Guégan J.F., Teriokhin A.T.* MultiTest V.1.2, a program to binomially combine independent tests and performance comparison with other related methods on proportional data // BMC Bioinformatics. 2009. 10: 443. doi:10.1186/1471-2105-10-443
3. *Fisher R.A.* Statistical Methods for Research Workers, 14th Edit. Edinburgh: Oliver and Boyd, 1970.
4. *Goudet J.* FSTAT (Version 1.2): A computer program to calculate F-statistics // Journal of Heredity. 1995. V. 86. P. 485–486.
5. *Goudet J.* An improved procedure for testing the effects of key innovations on rate of speciation // The American Naturalist. 1999. V.153. P. 549–555.
6. *Goudet J.* FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html. 2001. Updated from Goudet (1995).
7. *Holm S.* A simple sequentially rejective multiple test procedure // Scandinavian Journal of Statistics. 1979. V. 6. P. 65–70.
8. *Hugueny B., Guégan J.-F.* Community nestedness and the proper way to assess statistical significance by Monte Carlo tests: some comments on Worthen and Rohde's (1996) paper // Oikos. 1997. V. 80. P. 572–573.
9. *Manly B.F.J.* The Statistics of Natural Selection. London: Chapman & Hall, 1985.
10. *Prugnolle F., De Meeûs T., Durand P., Sire C., Théron A.* Sex-specific genetic structure in *Schistosoma mansoni*: evolutionary and epidemiological implications // Molecular Ecology. 2002. V.11. P. 1231–1238.

11. *Raymond M., Rousset F.* Genepop (Version-1.2) – Population-Genetics Software for Exact Tests and Ecumenicism // Journal of Heredity. 1995. V. 86. P. 248–249.
12. *Rice W.R.* Analyzing tables of statistical tests // Evolution. 1989. V. 43. P. 223–225.
13. *Rousset F.* GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux // Molecular Ecology Resources. 2008. V. 8. P. 103–106.
14. *Teriokhin A.T., De Meeûs T., Guégan J.F.* On the power of some binomial modifications of the Bonferroni multiple test // Zhurnal Obshchei Biologii. 2007. V. 68. P. 332–340.
15. *Whitlock M.C.* Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach // Journal of Evolutionary Biology. 2005. V. 18. P. 1368–1373.
16. *Wilkinson B.* A statistical consideration in psychological research // Psychological Bulletin. 1951. V. 48. P. 156–158.