Who is afraid of non-normal data? Choosing between parametric and non-parametric tests: a response

Rudolf Hoermann¹, John E M Midgley², Rolf Larisch¹ and Johannes W Dietrich^{3,4}

¹Department for Nuclear Medicine, Klinikum Lüdenscheld, Lüdenscheld, Germany, ²North Lakes Clinical, Ilkley, UK, ³Medical Department I, Endocrinology and Diabetology, Bergmannsheil University Hospitals, Ruhr University of Bochum, Bochum, Germany, and ⁴Ruhr Center for Rare Diseases (CeSER), Ruhr University of Bochum and Witten/Herdecke University, Bochum, Germany

Correspondence should be addressed to R Hoermann

Email

rudolf.hoermann@gmail.com

We read with great interest the Methodology Editorial on 'Choosing between parametric and non-parametric tests' by le Cessie *et al.* in the Journal (1). The topic is of wide interest to readers given that this is a frequent issue clinical researchers face in their daily work. The authors argue that the use of a *t*-test is preferable over non-parametric tests, as it does not necessarily require the data to be normally distributed, working well for moderately skewed distributions. They invoke the central limit theorem, which proves that the statistical mean in a larger sample ($n > 25-50$) is robustly independent of the normal distribution.

The question arises as to whether clinicians should be mainly interested in central tendency. We present an example for TSH where use of the *t*-test could be favoured following these recommendations despite the physiologically known non-normal distribution of the hormone (1). Two large clinical samples are compared – using some data from a previous prospective study (2) – a control group ($n=271$) where any contamination with thyroid autoimmune disease was carefully excluded and a test sample ($n=251$), which included untreated euthyroid (TSH within the reference range) subjects that tested positive for thyroid peroxidase antibodies (TPO Ab). TSH distribution in the samples is non-normal (Fig. 1A). The contamination of the test sample with autoimmune diseases is apparent in Fig. 1. Performing both parametric and non-parametric between-group tests returns the following statistical results, unpaired Welch *t*-test: mean difference 0.12 mIU/L (95%CI: $-0.04, 0.28$), $P=0.15$ and

Wilcoxon rank-sum test: $W=30,059$, $P=0.02$, using the R statistical package (version 3.6.2 for Mac (3)). The visibly observed difference between the two groups in these data is correctly indicated by Wilcoxon test, but not by the *t*-test.

For clinicians, it is important that existing group differences are reliably uncovered when performing statistical testing. For naturally skewed distributions like TSH in this example or BMI (4), the measures of central tendency may change only little when the distribution changes at the tail. The appropriate test must therefore aim to detect subtle differences between two groups, irrespectively whether they may or may not be accompanied by a shift in a single parameter of a distribution, usually the mean. Hence, we caution against the use of inappropriate means testing, for hormone parameters such as TSH.

This does not imply that testing of central tendencies will not succeed after careful inspection and possible further manipulation of the crude data, reducing distributional asymmetry (e.g. logarithmic transformation, outlier removal), not necessarily to achieve normality. However, issues remain with this approach. More extreme outcomes and change at the tail may potentially provide valuable information that should not be dismissed. Non-parametric tests or distributional comparisons may therefore offer more flexible tools to clinicians when dealing with skewed variables.

There are further caveats against relying on averaging and means comparison, particularly for TSH. Thyroid

Figure 1

(A) TSH distributions in two data samples, one excluding thyroid autoimmune disease ($n = 271$, broken line) and the other including TPO Ab - positive euthyroid thyroid autoimmune disease ($n = 251$, solid line). The distribution curve was derived by nonparametric kernel density estimation, using the R default function density with a Gaussian kernel and nrd0 bandwidth (3). For further statistical comparisons refer to text. Data for this demonstration were obtained during previous prospective sampling (2). (B) TSH distributions stratified by weight-adjusted LT4 dose, sampled during a recent retrospective study (8) across groups and visits in athyreotic patients with TSH concentrations within the reference range ($n = 425$). Differences between dosing groups are largely apparent at the tail, less so in their central tendencies. Density estimates differ significantly between groups (permutation test of equality (9)), group 1 vs 2 $P = 0.02$, 1 vs 3 $P < 0.001$, 2 vs 3 $P = 0.01$. The means for TSH in each

parameters display a high degree of individuality (low individuality index) in a population (5). The population or group mean plus its variance is therefore not shared among its members (6). Lacking a mathematical prerequisite for statistical averaging, these hormones cannot therefore be readily averaged in longitudinal studies (7). Doing so in violation of the ergodicity principle may result in misleading statistical outcomes (6). For example, the levothyroxine (LT4) dose effect on serum TSH levels during follow-up occurred largely at the tail of the distribution, rendering central tendencies and group averaging less informative (Fig. 1B).

Evaluation of change must take into account both the position of the person's mean relative to that of other group members and the temporal deviation of each person from their own mean. The reasons for the atypical behaviour of TSH are well known and physiologically rooted but beyond the scope of this brief statistical comment (10). However, in a controlling parameter where asymmetry and increased sensitivity at the tail are part of the physiological design, the primary clinical interest in subgroups or individuals does not concur with the analysis of central tendencies.

We conclude, statistical convenience testing should never be allowed to override physiological and mathematical principles, rather be guided by them, always taking into account the data structure and purpose of the statistical analysis.

Declaration of interest

J W D is co-owner of the intellectual property rights for the patent 'System and Method for Deriving Parameters for Homeostatic Feedback Control of an Individual' (Singapore Institute for Clinical Sciences, Biomedical Sciences Institutes, Application Number 201208940-5, WIPO number WO/2014/088516). The other authors have nothing to disclose.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

dose group, estimated with a generalised least-squares method (gls) and exchangeable correlation structure for subjects (3), are as follows, group 1 1.58 mIU/L (95% CI 1.38, 1.78), $n = 87$; group 2 1.38 mIU/L (95% CI 1.25, 1.52), $n = 197$ and group 3 ($n = 141$) 1.15 mIU/L (95% CI 1.00, 1.31), $n = 141$; p overall (ANOVA) = 0.003; P values (Tukey-Kramer test, unadjusted for multiple testing) for group 1 vs 2 $P = 0.10$, 1 vs 3 $P < 0.001$ and 2 vs 3 $P = 0.03$.

Author contribution statement

R H drafted the letter, J E M, R L, J W D contributed additional ideas and text to the final jointly approved letter.

References

- 1 Le Cessie S, Goeman JJ & Dekkers OM. Who is afraid of non-normal data? Choosing between parametric and non-parametric tests. *European Journal of Endocrinology* 2020 **182** E1–E3. (<https://doi.org/10.1530/EJE-19-0922>)
- 2 Hoermann R, Midgley JEM, Giacobino A, Eckl WA, Wahl HG, Dietrich JW & Larisch R. Homeostatic equilibria between free thyroid hormones and pituitary thyrotropin are modulated by various influences including age, body mass index and treatment. *Clinical Endocrinology* 2014 **81** 907–915. (<https://doi.org/10.1111/cen.12527>)
- 3 R Core Team. *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2020. (available at: <https://www.R-project.org/>)
- 4 Ejima K, Pavela G, Li P & Allison DB. Generalized lambda distribution for flexibly testing differences beyond the mean in the distribution of a dependent variable such as body mass index. *International Journal of Obesity* 2018 **42** 930–933. (<https://doi.org/10.1038/ijo.2017.262>)
- 5 Andersen S, Pedersen KM, Bruun NH & Laurberg P. Narrow individual variations in serum T4 and T3 in normal subjects: a clue to the understanding of subclinical thyroid disease. *Journal of Clinical Endocrinology and Metabolism* 2002 **87** 1068–1072. (<https://doi.org/10.1210/jcem.87.3.8165>)
- 6 Hoermann R, Midgley JEM, Larisch R & Dietrich JW. Functional and symptomatic individuality in the response to levothyroxine treatment. *Frontiers in Endocrinology* 2019 **10** 664. (<https://doi.org/10.3389/fendo.2019.00664>)
- 7 Molenaar PCM & Campbell CG. The new person-specific paradigm in psychology. *Current Directions in Psychological Science* 2009 **18** 112–117. (<https://doi.org/10.1111/j.1467-8721.2009.01619.x>)
- 8 Hoermann R, Midgley JEM, Dietrich JW & Larisch R. Dual control of pituitary thyroid stimulating hormone secretion by thyroxine and triiodothyronine in athyreotic patients. *Therapeutic Advances in Endocrinology and Metabolism* 2017 **8** 83–95. (<https://doi.org/10.1177/2042018817716401>)
- 9 Bowman AW & Azzalini A. R package ‘Sm’: nonparametric smoothing methods (version 2.2-5.6), 2018. (available at: <http://www.stats.gla.ac.uk/~adrian/sm>)
- 10 Hoermann R, Midgley JEM, Larisch R & Dietrich JW. Individualised requirements for optimum treatment of hypothyroidism: complex needs, limited options. *Drugs in Context* 2019 **8** 212597. (<https://doi.org/10.7573/dic.212597>)

Received 18 February 2020

Revised version received 20 April 2020

Accepted 29 April 2020