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

Saskia le Cessie[1,2], Jelle J Goeman[2] and Olaf M Dekkers[1]

Departments of [1]Clinical Epidemiology and [2]Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands

Correspondence should be addressed to S le Cessie
Email
S.le_Cessie@lumc.nl

Abstract

When statistically comparing outcomes between two groups, researchers have to decide whether to use parametric methods, such as the *t*-test, or non-parametric methods, like the Mann–Whitney test. In endocrinology, for example, many studies compare hormone levels between groups, or at different points in time. Many papers apply non-parametric tests to compare groups. We will explain that non-parametric tests have clear drawbacks in medical research, and, that's the good news, they are often not necessary.

European Journal of Endocrinology
(2020) **182**, E1–E3

In many papers the Methods' section reads like: 'for non-normally distributed data, non-parametric tests were used'. And indeed, many papers apply non-parametric tests, such as Mann–Whitney test or Wilcoxon test, to compare groups, when the data do not seem completely normally distributed. However, the use of parametric methods, like the *t*-test, has a clear advantage compared to non-parametric tests: where a non-parametric test will only produce a *P* value, a *t*-test will also produce the observed mean difference between the groups, with a 95% confidence interval (CI). For example, a mean difference in TSH between groups of 0.35 mU/L, with a 95% CI from 0.12 to 0.58 mU/L. This is useful and important information: it shows the size and direction of the observed effect, with the precision of the estimated difference; such information is crucial to determine whether the results are clinically relevant. In contrast, a non-parametric test only provides a *P* value, a quantity that is often misinterpreted and that cannot be used to judge the clinical relevance of difference (1).

In observational research, groups are often not directly comparable. For example, it may be that the above-mentioned difference in TSH is partly confounded by age. In such situations researchers should perform adjusted statistical analyses. Here is a second advantage of parametric methods; they have a direct link with regression models, which enables the researcher to provide an effect estimate (for example, difference in TSH) that is adjusted for other variables which differ between the groups (for example age). In fact, a simple *t*-test will yield identically the same results as an unadjusted linear regression model. Therefore, it is consistent to use a *t*-test to compare certain outcomes between groups, if a linear regression model would be used for the same outcomes when adjusting for confounding is deemed necessary. For linear regression models the same assumptions hold as for *t*-tests.

However, many researchers believe that *t*-tests may only be used when the outcome variable is normally distributed. Fortunately, this is not true. The *t*-test is not afraid of non-normal data. When there are more than about 25 observations per group and no extreme outliers, the *t*-test works well even for moderately skewed distributions of the outcome variable.

https://eje.bioscientifica.com
https://doi.org/10.1530/EJE-19-0922

Published by Bioscientifica Ltd.

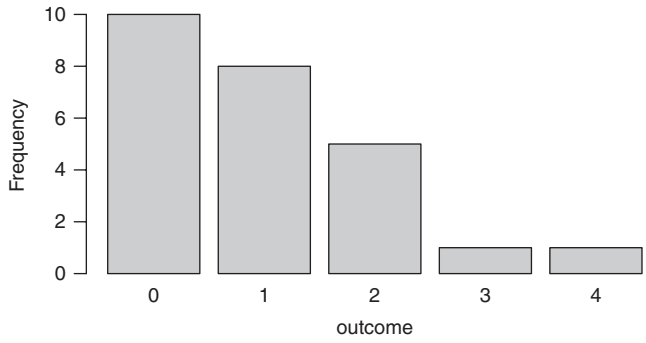

Figure 1
A moderately skewed distributed outcome with 25 observations.

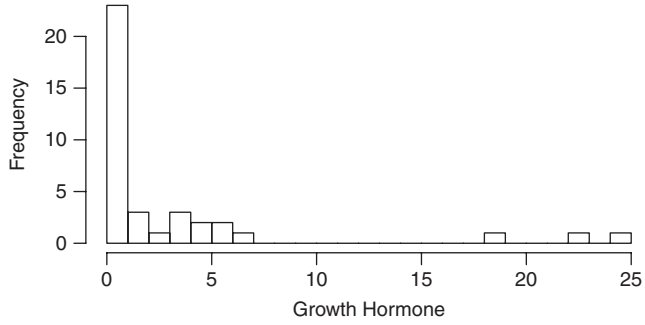

Figure 2
A very skewed distributed outcome, where a logarithmic transformation of the outcome should be performed before using a *t*-test.

Consider a distribution of the outcome in 25 patients given in Fig. 1. Most researchers would consider these data non-normally distributed and therefore apply a non-parametric test. Although these data are indeed not normally distributed, the *t*-test will work fine in this situation. The same holds in many situations where the distribution is not completely normal: the *t*-test and thus also linear regression models perform statistically well. The larger the sample size, the more extreme the distribution of the observations can be without compromising the validity of the *t*-test (This is because of the central limit theorem, which states that the distribution of the mean will approximate to the normal distribution when the sample size increases, regardless of the distribution of the original observations (under some regularity conditions).). Most hormones, for example TSH and prolactin, have distributions that allow to use *t*-tests with moderate sample sizes of 25 patients or more.

Some variables are by nature extremely skewed, such as CRP levels or growth hormone concentrations (see Fig. 2 for an example). Here, the few very large outcomes will strongly influence the results of a *t*-test. In that case a transformation of the data can be performed to obtain a more normal distribution and more stable results. For positive data with some extreme high values (growth hormone in acromegaly or prolactin values in prolactinomas for example) logarithmic transformations are commonly used, because results then have a relatively straightforward interpretation: for example, after performing a 2-log transformation, a difference of 1 unit on the log scale means that the mean in one group is twice as large as the mean in the other group.

When groups sizes are small (as a rule of thumb: below 25), the outcome variable should be normally distributed to use the *t*-test. It is difficult to see from observed data whether they are sufficiently normal, because visual tools to detect normality, such as histograms, are not very informative in small samples. Knowledge from other studies can be used to decide if normality may be assumed. If so, the *t*-test can still be used. Otherwise, a non-parametric statistical test is preferred. For example, variables as height and blood pressure can be assumed to be normally distributed without looking at the data.

Some statisticians advise to perform a statistical test for normality, such as the Shapiro–Wilk test, and to let the decision between parametric and non-parametric methods depend on the significance of that test. We disagree with this approach. In small samples, where normality is an important assumption, tests for normality do not have much power, while for larger samples, where deviation of normality is no longer an obstacle to use a *t*-test, the tests for normality will often be statistically significant. The test for normality will therefore often suggest to use the wrong test. For example, performing a Shapiro–Wilk test for normality on the data of Fig. 1 will yield a *P* value of 0.0007. However, because there are 25 observations and no extreme outliers, the *t*-test will yield valid results here.

In the Table 1 of a paper, non-normally distributed variables are commonly described with medians and interquartile ranges. Still, as argued, when group sizes are large enough, it is both perfectly allowed as well as informative to compare the groups with a *t*-test and report the mean difference with 95% confidence interval. To summarize, *t*-tests can often be used to compare continuous variables between groups, even if the underlying distribution of the observations is not normal. The main advantage of parametric methods such as *t*-tests

or linear regression is that they provide effect estimates, which allows researcher to examine the clinical relevance of the results.

Declaration of interest

Olaf M Dekkers is a Deputy Editor for *European Journal of Endocrinology*. He was not involved in the review or editorial process for this paper, on which he is listed as an author. 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.

Reference

1 Dekkers, OM. Why not to (over)emphasize statistical significance. *European Journal of Endocrinology* 2019 **181** E1–E2. (https://doi.org/10.1530/EJE-19-0531)

Received 13 November 2019
Accepted 21 November 2019