Editorial

The Limitation of Widely Used Data Normality Tests in Clinical Research

Mohd Normani Zakaria

Audiology Programme, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Malaysia

Use your device to scan and read the article online

Citation: Zakaria MN. The Limitation of Widely Used Data Normality Tests in Clinical Research. Aud Vestib Res. 2022;31(1):1-3.

https://doi.org/10.18502/avr.v31i1.8127

Highlights

- It is imperative to analyse the data distribution
- The widely used data normality tests have a limitation
- Graphical presentations of data distribution should be considered

Corresponding Author:
Audiology Programme, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Malaysia
mdnorman@usm.my

In clinical research, statistical analysis plays a prominent role in achieving the desired study outcomes. Depending on the specific study objectives, appropriate statistical tests are employed to decide whether to accept or reject the null hypothesis so that a concrete conclusion can be made. Herein, parametric and/or non-parametric tests are selected to analyse the research data. As reported elsewhere, the use of parametric tests is superior to non-parametric analyses due to their higher power in rejecting null hypotheses [1, 2].

It is well known that the data distribution must be checked prior to the application of any statistical tests. Parametric tests can only be applied if the data are normally distributed (a bell-shaped curve), whereas non-normal (skewed) data distribution necessitates the use of non-parametric statistical analyses. There are at least two ways to assess the data distribution: the visual inspection of the data and the use of inferential normality tests. To visually inspect the research data, histograms, Q-Q plots, and others can be used. The limitation of this approach is that it is highly subjective. On the other hand, to "objectively" decide whether the data are normally distributed (or not), normality tests such as Kolmogorov-Smirnov, Shapiro-Wilk, and Chi-square tests are commonly conducted. This "p-value" approach seems simpler and easier to report (e.g. if the p value is greater than 0.05, the data are considered to have a normal distribution). It is indeed more popular among many researchers (relative to the visual inspection) in deciding the appropriate statistical tests.

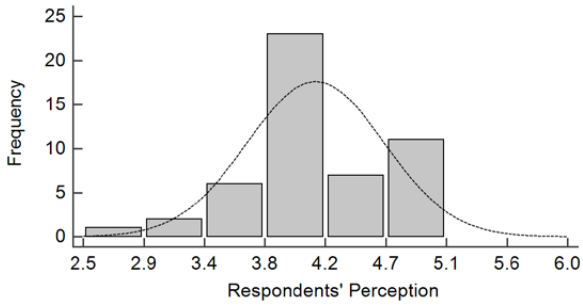

Figure 1. Histogram of data distribution for a respective item on the questionnaire (n=50)

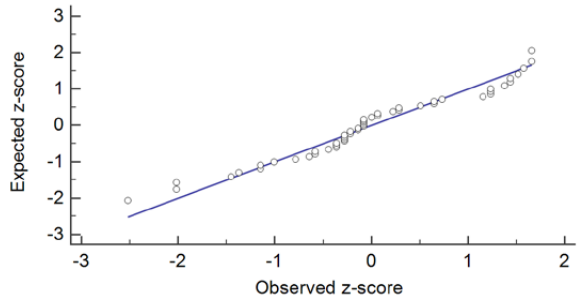

Figure 2. Q-Q plot of data distribution for a respective item on the questionnaire (n=50)

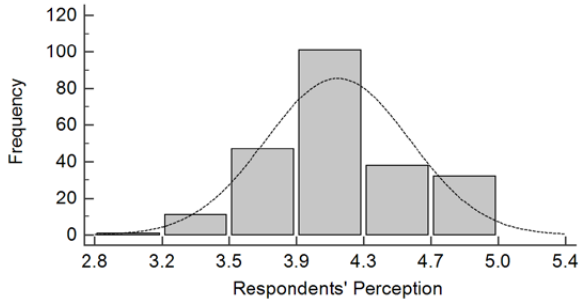

Figure 3. Histogram of data distribution for a respective item on the questionnaire (n=230)

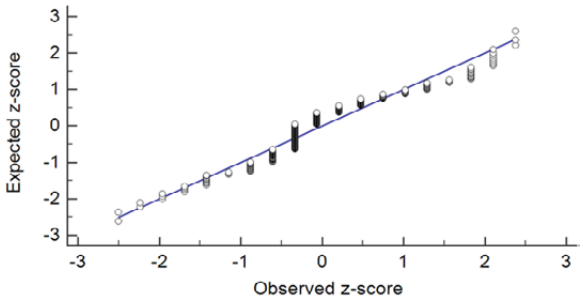

Figure 4. Q-Q plot of data distribution for a respective item on the questionnaire (n=230)

Nevertheless, it is worth noting that questionable results can be obtained with the application of the aforementioned normality tests, particularly when the sample size is large. To demonstrate this, data from a study by Ee [3] titled "Personal listening devices: a survey of attitudes, knowledge, and usage patterns among students at the School of Health Sciences, Universiti Sains Malaysia" are used. In this study, a dedicated questionnaire is employed to compare the response to specific items on the questionnaire between male and female students. In this paper, one specific item is chosen as an example, i.e. "Listening to music while doing assignments or studying helps me concentrate" and the response option is "strongly disagree" (1), "disagree" (2), "unsure" (3), "agree" (4) and "strongly agree" (5).

Figure 1 shows the histogram that plots the data distribution (n=50), and the respective Q-Q plot is shown in Figure 2. As illustrated, the data appear to be normally distributed (with no notable difference between the tallest bar of the histogram and the bell-shaped normality curve). The Q-Q plot is consistent with the result of histogram (i.e., most of the dots do not stray from the straight line) (Figure 2). Table 1 shows the statistical results of several normality tests including D'Agostino-Pearson test (that is less commonly used in research). As indicated, all the statistical results agree with each other ($p>0.05$), indicating that the data are normally distributed. Subsequently, an independent t-test (parametric analysis) is conducted to compare the response between males (n=25, mean=4.12, SD=0.57) and females (n=25, mean=4.24, SD=0.40). The resultant p value is 0.377, that suggests no significant difference in the perception between genders for the respective item (null hypothesis is accepted).

On the other hand, when the number of students participating in the study is increased (n=230), different outcomes are observed. As shown in Table 1, except

Table 1. Statistical results (p values) of different normality tests for different sample sizes

	Normality test			
Sample size	Kolmogorov-Smirnov	Shapiro-Wilk	Chi-square	D'Agostino-Pearson
50	0.100	0.925	0.064	0.798
230	<0.001	<0.001	<0.001	0.099

for D'Agostino-Pearson test (p=0.099), all the normality test results are significant (p<0.05), implying that the data are not normally distributed. In view of this, a non-parametric Mann Whitney test is carried out and the resultant p value is 0.054. Based on this value, the null hypothesis is accepted suggesting that the perception for the item "Listening to music while doing assignments or studying helps me concentrate" is comparable between male and female students.

Interestingly, the visual inspection approach reveals more noteworthy outcomes. Figure 3 shows the histogram of the data distribution (n=230). As clearly illustrated, the data look "very much" normally distributed (and the tallest bar of the histogram is even closer to the normality curve, compared to that in Figure 1). The Q-Q plot (Figure 4) is consistent with the respective histogram, supporting the normality of the data distribution. Based on the outcomes of the visual inspection method, it is clear that the parametric test can be used (as there is no evidence of skewed distribution). In fact, when the response to the respective item is compared between males (n=105, mean=4.09, SD=0.45) and females (n=125, mean=4.19, SD=0.35) by means of the independent t-test, the p value is now significant (0.046). In this regard, the null hypothesis is rejected, and a better study conclusion can be made.

Generally, obtaining "positive" outcomes is the ultimate aim of any research. In this regard, rejecting the null hypothesis can be considered a favourable outcome by many researchers. The possibility of rejecting the null hypothesis increases by employing appropriate statistical tests along with larger sample sizes. This is because the "p value" is highly dependent on the sample size [4]. By increasing the sample size, the power of the study may increase and the likelihood of rejecting the null hypothesis increases (i.e. p<0.05). In the context of normality testing, increasing the sample size may also result in the rejection of the null hypothesis, implying that the tested data are "different" from the normal distribution. This scenario may not be seen if the sample size is smaller. On the other hand, histograms and Q-Q plots are more conservative and realistic, regardless of the sample size, which can be advantageous in making the "right" decisions on the data distribution.

Taken together, the inferential normality tests are still beneficial and serve their purpose when reporting research outcomes. Nevertheless, researchers should be aware that these tests can be "overpowered" when the sample is large, leading to undesired outcomes. Herein, it is good to include the graphical presentations of data distribution and collectively, a more accurate decision can be made so that appropriate statistical tests can be chosen. Lastly, the D'Agostino-Pearson test appears to be less affected by the sample size and can be a good alternative when analysing the data distribution.

References

[1] Tanizaki H. Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. J Appl Stat. 1997;24(5):603-32. [DOI:10.1080/02664769723576]

[2] Nahm FS. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean J Anesthesiol. 2016;69(1):8-14. [DOI:10.4097/kjae.2016.69.1.8]

[3] Ee PW. Personal listening devices: a survey of attitudes, knowledge, and usage patterns among students at the School of Health Sciences, Universiti Sains Malaysia. [Undergraduate thesis]. Kelantan, Malaysia: Universiti Sains Malaysia; 2018.

[4] Zakaria MN. The values of effect size in statistical decision for clinical research. Aud Vestib Res. 2017;26(1):1-3.