How do normality tests behave for rounded data?

Marina Iturrate-Bobes[1] **Ignacio Montes**[1] **Raúl Pérez-Fernández**[1]

[1]Department of Statistics and O.R. and Mathematics Didactics, University of Oviedo, Spain

Most statistical inference techniques are built assuming the existence of a continuous random variable that should be measured repeatedly. However, there are situations in which the imprecision inherent to the data collection process affects the results. For instance, this occurs in the presence of censored data, missing data, or when the data is rounded. In all these cases, the results of statistical tests might be compromised. As one such example, we tackle the problem considered in [4, 5] of testing for normality with rounded data, considering here the perspective of imprecise probabilities. More specifically, we provide a theoretical framework and perform an empirical study for such a problem. For this aim, we take the following steps:

Step 1-Mathematical formalisation: We formalise the problem using random sets [1]. Assuming that the continuous random variable of interest is rounded to the d-th decimal number ($d \in \mathbb{Z}$), where $d \leq 0$ means rounding to the $(1-d)$-th digit to the left of the decimal point, we define the set $\mathbb{Z}_d = \{x \mid 10^d \cdot x \in \mathbb{Z}\}$. Hence, we consider a random set $\tilde{X}$ such that for any element ω in the possibility space it holds that

$$\tilde{X}(\omega) = \left[z - 5 \cdot 10^{-(d+1)}, z + 5 \cdot 10^{-(d+1)}\right)$$

for some $z \in \mathbb{Z}_d$. For example, if data is rounded to the second decimal number (i.e., $d = 2$) and for some ω it holds that $X(\omega) = 1.7428$, we will observe the rounded value 1.74. This rounded value might appear after rounding any value in the interval $\tilde{X}(\omega) = [1.735, 1.745)$. In this framework, all we know about the "real" rounded random variable X is that it is one of the measurable selections of $\tilde{X}$.

Step 2-Tests under imprecision: In statistical testing, a hypothesis test determines the acceptance region and the critical region. However, when the data is subject to imprecision, according to [2] a third region, called the indecision region, naturally appears, and it is formed by those random samples for which the imprecision prevents from making a decision.

Step 3-Normality tests: Many normality tests can be found in the literature (e.g., Shapiro-Wilk, Lilliefors, Cramer-von Mises, Anderson-Darling, Jarque-Bera, etc.). For these tests, the power does not depend on the mean or the standard deviation of the population. In our setting, we analyse whether, when applied to rounded data, the tests become sensitive to these population parameters.

Step 4-Empirical analysis of the power: An empirical analysis is performed to (i) examine how the power varies when varying the decimal number at which the data is rounded, and (ii) compare the performance of different normality tests in the presence of imprecision.

After following these steps, we will show how random sets allow us to formalise the imprecision caused by rounding in order to have a full picture of how normality tests behave under such imprecision. For future research, we will explore the approach to goodness-of-fit tests proposed in [3], where the Kolmogorov-Smirnov test was expanded to interval-data by using p-boxes. More precisely, we will look for a link between the approach based on random sets and the approach based on p-boxes when testing for normality with rounded data.

References

[1] I. Couso, D. Dubois, and L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Springer Briefs in Applied Sciences and Technology. Springer, 2014. DOI: 10.1007/978-3-319-08611-8.

[2] I. Couso and L. Sánchez. "Mark-recapture techniques in statistical tests for imprecise data". In: *International Journal of Approximate Reasoning* 52 (2011), pp. 240–260. DOI: 10.1016/j.ijar.2010.07.009.

[3] S. Destercke and O. Strauss. "Kolmogorov-Smirnov Test for Interval Data". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Ed. by A. Laurent et al. Springer, 2014, pp. 416–425. DOI: 10.1007/978-3-319-08852-5_43.

[4] A.R. Tricker. "The effect of rounding on the power level of certain normal test statistics". In: *Journal of Applied Statistics* 17.2 (1990), pp. 219–228. DOI: 10.1080/757582833.

[5] A.R. Tricker. "The effect of rounding on the significance level of certain normal test statistics". In: *Journal of Applied Statistics* 17.1 (1990), pp. 31–38. DOI: 10.1080/757582644.