How do normality tests behave for rounded data?

Marina *Iturrate-Bobes* iturratemarina@uniovi.es
Ignacio *Montes* imontes@uniovi.es
Raúl *Pérez-Fernández* perezfernandez@uniovi.es

Funded by project PID2022-140585NB-I00

1 Introduction to normality tests

Why normality tests?

- They are statistical procedures used to determine whether some data come from a normal distribution.
- Their study is fundamental as the assumption of normality is a necessary condition for certain inferential tests such as the t-test and the ANOVA.

Formally...

$$\begin{cases} H_0 : X \equiv N(\mu, \sigma) \\ H_1 : X \not\equiv N(\mu, \sigma) \end{cases}$$

for certain $\mu = E(X) \in \mathbb{R}$, $\sigma^2 = Var(X) \in \mathbb{R}^+$.

Common normality tests

- Anderson-Darling [1].
- Lilliefors [4].
- Shapiro-Wilk [5].

Property: distribution-free within any location-scale family

$$\text{p-value}(\mathbf{x}) = \text{p-value}(a\mathbf{x} + b)\,,$$

for a sample $\mathbf{x}$ and any $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$.

2 The problem with rounded data

The problem

Unfortunately, when collecting data from a continuous random variable there is an inevitable error caused by rounding that might compromise the results of the statistical tests.

Formally...

- The continuous random variable X is rounded to the d-th decimal number ($d \in \mathbb{Z}$).
- $d \leq 0$ means rounding to the $(1-d)$-th digit to the left of the decimal point.
- Set of possible values: $\mathbb{Z}_d = \{x \mid 10^d \cdot x \in \mathbb{Z}\}$.

Behaviour

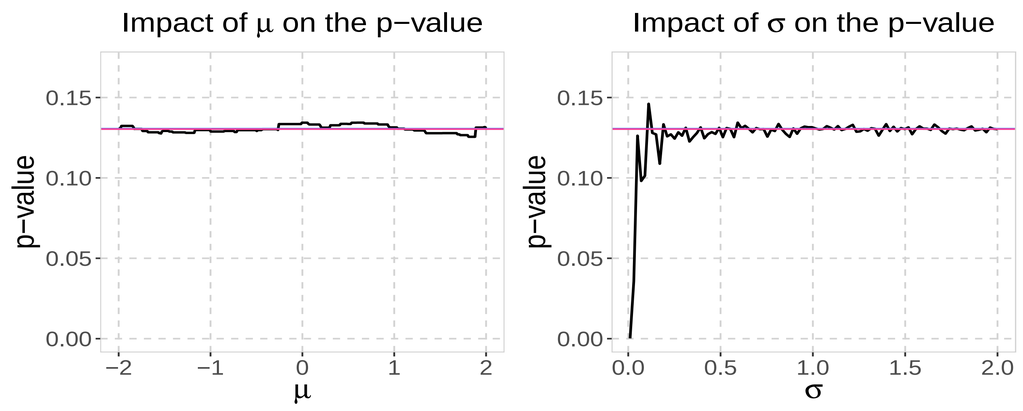
- For any $n \in \mathbb{N}$, the rounding function to the d-th decimal number is $\varphi : \mathbb{R}^n \to (\mathbb{Z}_d)^n$.
- For any $b \in \mathbb{R}$, $\varphi(\mathbf{x} + b) \approx \varphi(\mathbf{x}) + b$.
- Unfortunately, multiplying by a small positive constant $a \in \mathbb{R}^+$ before rounding ($\varphi(a\mathbf{x})$) reduces the granularity of the result compared to rounding first and then multiplying ($a\varphi(\mathbf{x})$). For large positive constants $a \in \mathbb{R}^+$, the opposite effect occurs although it is relatively less significant. Therefore:

$$\text{p-value}\big(\varphi(a\mathbf{x} + b)\big) \approx \text{p-value}\big(\varphi(a\mathbf{x}) + b\big)\big) = \text{p-value}\big(\varphi(a\mathbf{x})\big)\,,$$

which may differ greatly for small values of $a \in \mathbb{R}^+$ from

$$\text{p-value}\big(a\varphi(\mathbf{x})\big) = \text{p-value}\big(\varphi(x)\big)\,.$$

Example: Small impact of μ (left) and potentially-large impact of σ (right) on the p-value of the Shapiro-Wilk test for a sample $\sigma\mathbf{x} + \mu$ from $X \equiv N(0,1)$ rounded to the third decimal number.



3 A random set-based approach

The model

Random sets [2] can be used for modelling the rounding problem: when we observe a rounded random variable $\tilde{X}$ taking values on $\mathbb{Z}_d$, we may construct the random (interval) set:

$$\Gamma = [\tilde{X} - 5 \cdot 10^{-(d+1)}, \tilde{X} + 5 \cdot 10^{-(d+1)}[\,.$$

Still, we are interested in the unobservable random variable X, of which we only know that $\tilde{X} = \varphi(X)$ and that X is one of the measurable selections of Γ.

The decision

When extending statistical inference from random variables to random sets [3], the p-value associated with an interval-valued sample $I_1, \ldots, I_n$ becomes a set:

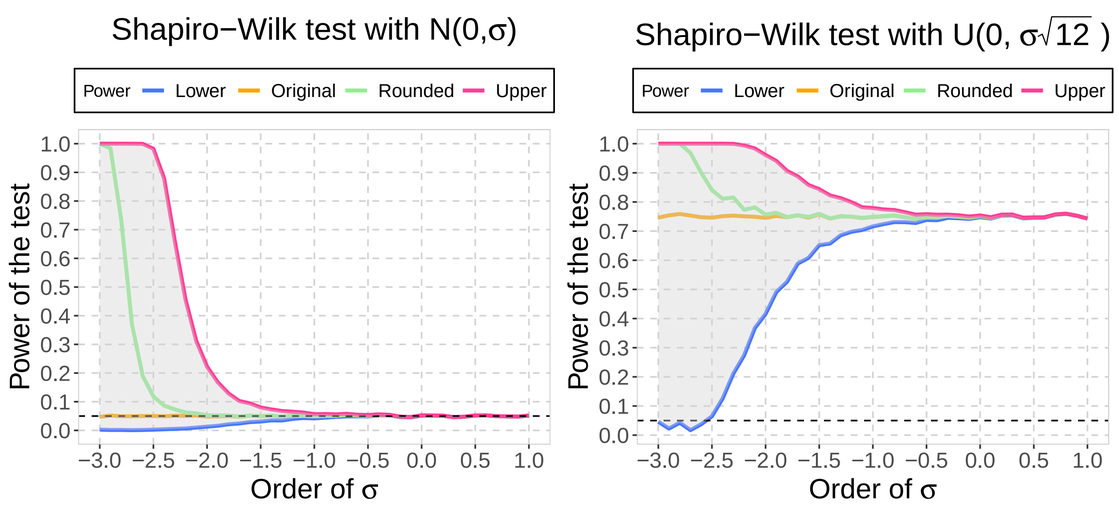
$$\text{p-value}(I_1, \ldots, I_n) = \{\text{p-value}(x_1, \ldots, x_n) \mid x_i \in I_i\,,\ \forall i \in \{1, \ldots, n\}\}\,.$$

A three-way partition of the (interval-valued) sample space arises:

1. If $\sup$ p-value$(I_1, \ldots, I_n) \leq \alpha$, then $I_1, \ldots, I_n$ belongs to the Rejection Region (RR), and H_0 is rejected.
2. If $\inf$ p-value$(I_1, \ldots, I_n) > \alpha$, then $I_1, \ldots, I_n$ belongs to the Acceptance Region (AR), and H_0 is not rejected.
3. If $\inf$ p-value$(I_1, \ldots, I_n) \leq \alpha < \sup$ p-value$(I_1, \ldots, I_n)$, then $I_1, \ldots, I_n$ belongs to the Indecision Region (IR), and no decision can be made due to the effect of not knowing the exact values of the sample (i.e., due to the effect of imprecision).

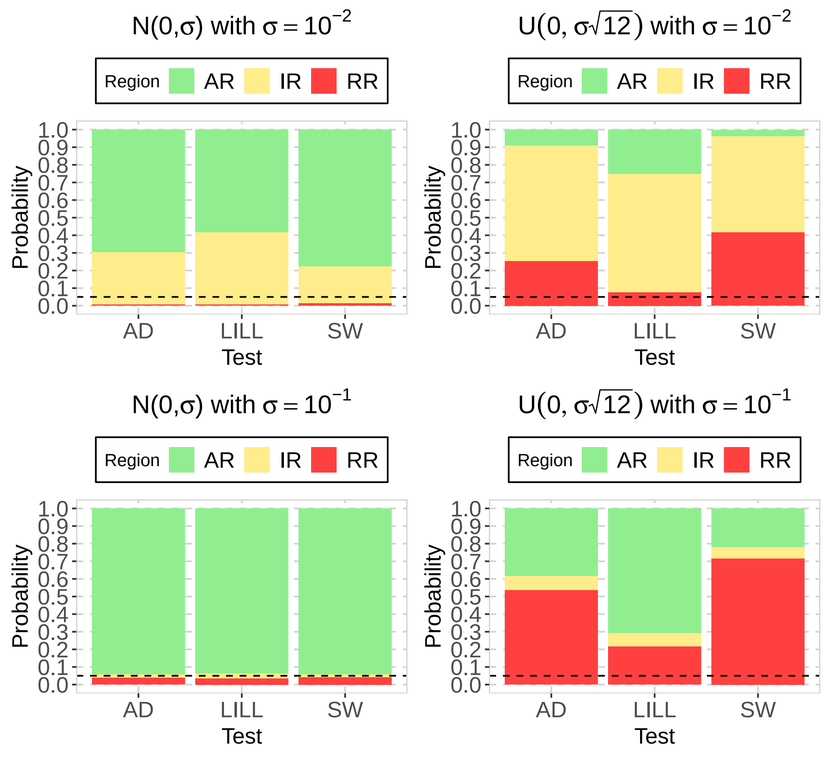
4 Results of the experimental setup

The figures below represent the power of the Shapiro-Wilk normality test estimated by using Monte Carlo simulation with 10^4 repetitions for samples of size $n = 50$ from $X \equiv N(0, \sigma)$ (left) and $X \equiv U(0, \sigma\sqrt{12})$ (right), with $\sigma \in]10^{-3}, 1[$, which were rounded to the third decimal number.



Applying the test directly to the rounded samples leads to powers of 1 for small values of σ (green line), even under the null hypothesis. Additionally, the probability of the Indecision Region (difference between the pink and blue lines) decreases as σ increases for both distributions.

Similar results are obtained when comparing the probabilities of the three regions for Anderson-Darling (AD), Lilliefors (LILL) and Shapiro-Wilk (SW) for fixed σ.



5 Conclusions

- We have explored the use of normality tests for rounded data.
- Ignoring the effect of rounding might lead to erroneous conclusions.
- The impact of μ is irrelevant... but small values of σ may cause a great impact on the p-value.
- We propose addressing the problem from the perspective of random sets.
- The Shapiro-Wilk test leads to a smaller IR under the null hypothesis and to a greater RR under the alternative hypothesis of uniformity.

6 References

[1] T.W. Anderson and D.A. Darling. "A test of goodness of fit". In: Journal of the American Statistical Association 49.268 (1954), pp. 765–769.

[2] 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.

[3] 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.

[4] H.W. Lilliefors. "On the Kolmogorov–Smirnov test for normality with mean and variance unknown". In: Journal of the American Statistical Association 62.318 (1967), pp. 399–402.

[5] S.S. Shapiro and M.B. Wilk. "An analysis of variance test for normality (complete samples)". In: Biometrika 52.3-4 (1965), pp. 591–611.