Assesing five normality tests: Lilliefors, Cramér-von Mises, Anderson-Darling, Shapiro-Wilk and Shapiro-Francia

Diego Ramos Crespo

May 7, 2025

1 Introduction

Assessing whether data follows a normal distribution is crucial for various statistical applications. Over the years, several tests have been developed for this purpose. This document evaluates these tests to identify which one is most sensitive in detecting deviations from normality. The tests considered are the Lilliefors Test (LT), Cramér-von Mises Test (CVMT), Anderson-Darling Test (ADT), Shapiro-Wilk Test (SWT), and Shapiro-Francia Test (SFT).

The data were randomly generated in R using the `rnorm()` function, with sample sizes of 60, 240, and 960. The sample was then contaminated by randomly inserting data values from a t-Student distribution with $\nu = 12$ degrees of freedom at different contamination levels: 0%, 10%, 20%, and 30%.

2 Simulation

The performance of the previous five normality tests was evaluated under different contamination levels and sample sizes. The experiment was repeated for three different sample sizes: $n = \{60, 240, 960\}$, $n = 240$. For each combination of sample size and contamination level, five normality tests were applied: Lilliefors (LT), Cramér–von Mises (CVMT), Anderson–Darling (ADT), Shapiro–Wilk (SWT), and Shapiro–Francia (SFT). Each test was evaluated at a significance level of $\alpha = 0.10$. The entire process was repeated $B = 10^4$ and $B = 10^3$ times, and the proportion of rejections of the null hypothesis of normality was recorded for each case.

Figure 1 shows the proportion of rejections obtained for each normality test, depending the contamination level and sample size. In general, for all methods, the rejection rate increases as the contamination level increases, particularly for larger sample sizes.

LT (Lilliefors Test): The table on the left shows the rejection proportions for the Lilliefors test (LT) at a significance level of $\alpha = 0.10$. It can be observed that for each sample size ($n = 60, 240, 960$), the rejection proportion tends to increase as the contamination level rises (from 0% to 30%). This indicates that the test gains statistical power to detect deviations from normality when outliers are present, especially with larger sample sizes such as $n = 960$.

CVMT (Cramér–von Mises Test): The table on the right presents the results for the Cramér–von Mises test (CVMT). Similar to LT, the rejection proportions increase progressively as the contam-

ination level increases. The growth is more pronounced for larger samples, suggesting that CVMT is also sensitive to contamination and that its statistical power improves with greater sample size.

ADT (Anderson–Darling Test): The results for the Anderson–Darling test (ADT) are shown in the table on the left. This test shows a clear increasing trend in rejection proportions as the contamination level rises, especially in medium and large sample sizes. For $n = 960$ and 30% contamination, the rejection rate reaches 19.41%, which indicates a strong ability to detect non-normality under contaminated conditions.

SWT (Shapiro–Wilk Test): The table on the right shows the results for the Shapiro–Wilk test (SWT). This test stands out for its high sensitivity to contamination: for a sample size of $n = 960$, the rejection proportion exceeds 36% when contamination reaches 30%. Even in smaller samples, the increase is significant, suggesting that SWT is particularly effective in detecting non-normality caused by outliers.

SFT (Shapiro–Francia Test): Finally, the table corresponding to the Shapiro–Francia test (SFT) shows a pattern similar to the Shapiro–Wilk test. As contamination increases, the rejection proportion also rises, reaching 43.37% for $n = 960$ with 30% contamination. This suggests that SFT is highly sensitive to the presence of anomalous data and therefore useful in scenarios where moderate to severe deviations from normality are suspected.

In particular, the Shapiro–Wilk (SWT) and Shapiro–Francia (SFT) tests show higher sensitivity to contamination; this can be observed by the significant increase in rejection rates when moving from 0% to 30% contamination reaching values up to 0.3689 and 0.4337 for the largest sample size $n = 960$ respectivey. Meanwhile Anderson-Darling test seems to remain pretty similar even with the increase of contamination. This behavoir is similiar compared with tests like Lilliefors (LT) or Cramér–von Mises (CVMT), which exhibit a more moderate increase under the same conditions.

LT (Lilliefors)

% \ n	60	240	960
0	0.0990	0.1068	0.1006
0.1	0.1026	0.1018	0.1126
0.2	0.1088	0.1152	0.1299
0.3	0.1081	0.1221	0.1499

CVMT (Cramér–von Mises)

% \ n	60	240	960
0	0.1012	0.0998	0.0996
0.1	0.1016	0.0984	0.1101
0.2	0.1119	0.1151	0.1436
0.3	0.1152	0.1279	0.1732

ADT (Anderson–Darling)

% \ n	60	240	960
0	0.0983	0.0981	0.0984
0.1	0.1039	0.1025	0.1151
0.2	0.1156	0.1229	0.1534
0.3	0.1234	0.1408	0.1941

SWT (Shapiro–Wilk)

% \ n	60	240	960
0	0.0982	0.0989	0.1023
0.1	0.1141	0.1301	0.1828
0.2	0.1280	0.1669	0.2862
0.3	0.1429	0.2027	0.3689

SFT (Shapiro–Francia)

% \ n	60	240	960
0	0.1003	0.1009	0.1057
0.1	0.1206	0.1467	0.2182
0.2	0.1414	0.2002	0.3409
0.3	0.1576	0.2441	0.4337

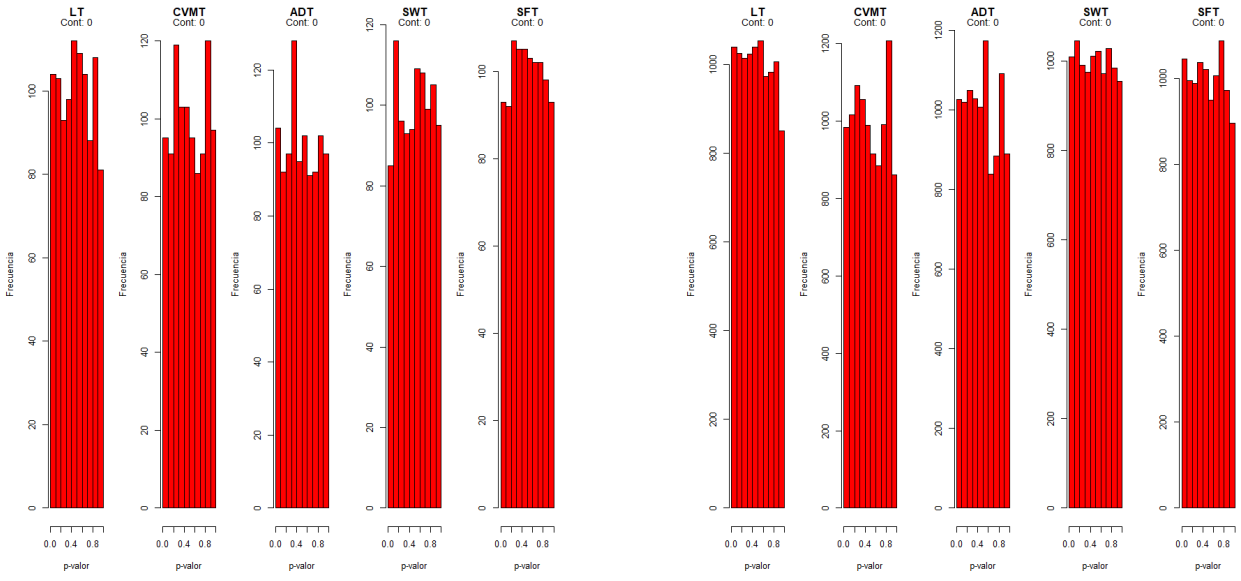
Figure 1: Proporción de rechazos ($\alpha = 0.10$) para cinco pruebas de normalidad bajo diferentes niveles de contaminación y tamaños muestrales.

In summary, the analysis of these tables suggests that the power of normality tests varies significantly across methods and is strongly affected by both the sample size and the level of contamination.

3 p-value distributions

For every test, the null hypothesis (H_0) assumes that the data is drawn from a normal distribution. In case the data is contaminated with data that comes from a different distribution, then the tests begin to detect these deviations, and p-values become increasingly concentrated near 0, reflecting stronger evidence against the null hypothesis. After simulating $1,000$ and $10,000$ p-values for each of the tests, it was possible to plot histograms with 10 bins each. It is important to highlight that the computation time was very long for simulating the samples of size 10^4, therefore, we also decided to try the simulation with samples of size 10^3 in order to compare both performances.

Figure 2 shows the results obtained using each method. In this first image, it can be observed the control group, this means that the normally distributed sample is not contaminated. This plot can help us understand how does the test ideally must work. For comparative purposes, we decided to plot histograms for both sample sizes. In subfigure on the left (2a), it is displayed the histogram generated using 10^3 iterations. Meanwhile, in subfigure on the right (2b) were used 10^4 iterations, however, the results seems not be extremely different, both show a pretty similar behavior. Thus, the expected result is to obtain uniformly distributed p-values when conducting an experiment to assess wether data follows a normal distribution or not.



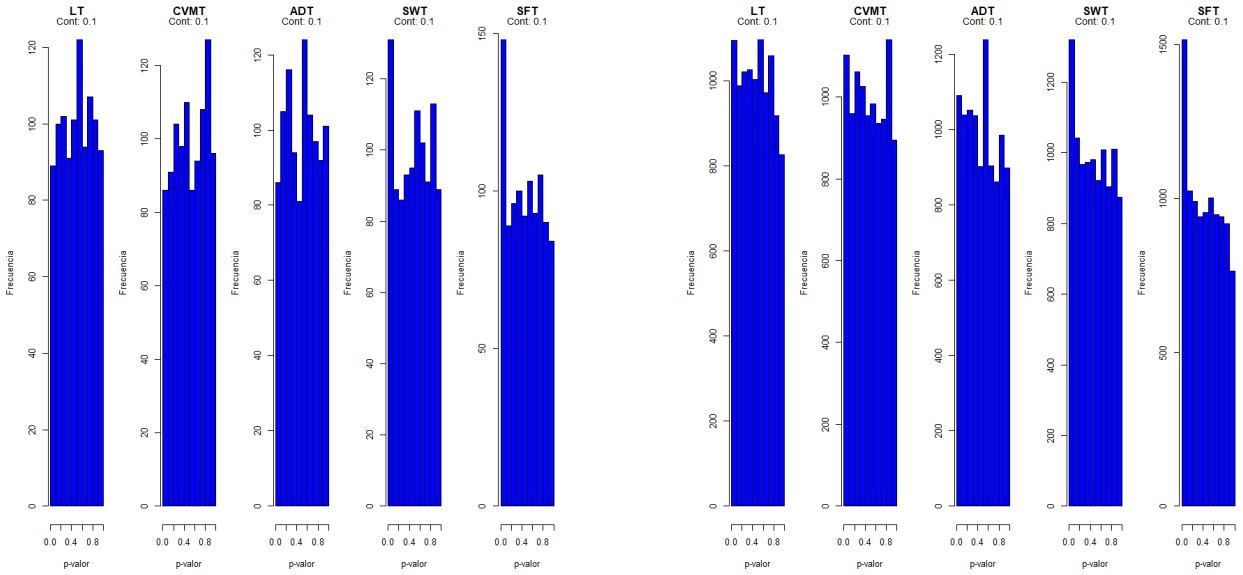
(a) Sample size (n) = 240. Iterations = 1,000

(b) Sample size (n) = 240. Iterations = 10,000

Figure 2: Histograms of p-values under 0% contamination levels.

In Figure 3, a 10% contamination level has been introduced by replacing part of the standard normal sample with observations from a Student-t distribution with ν =12 degrees of freedom. As

seen in both subfigures, 10^3 iterations on the left (3a) and 10^4 on the right (3b), the distribution of p-values starts to deviate from uniformity. There is a noticeable increase in the frequency of lower p-values, indicating that the normality tests begin to detect mild deviations from the normal distribution. This marks the onset of decreasing p-values as contamination increases.

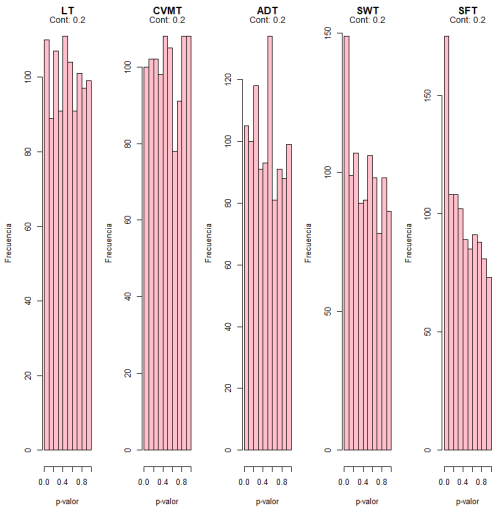


(a) Sample size (n) = 240. Iterations = 1,000

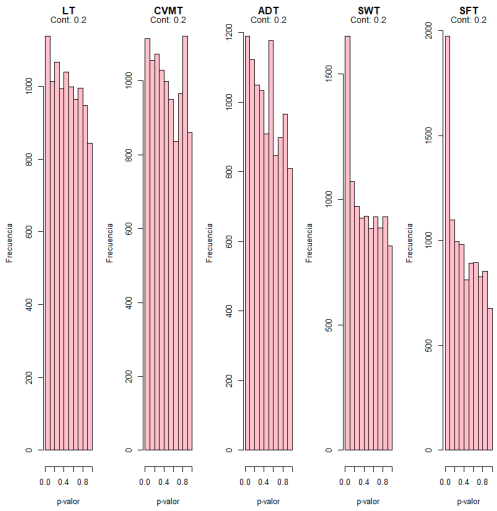
(b) Sample size (n) = 240. Iterations = 1,000

Figure 3: Histograms of p-values under 10% contamination level.

Finally, Figure 4 presents the p-value distributions under a 20% contamination level. The left subfigure (4a) with 10^3 iterations and the right subfigure (4b) with 10^4 iterations both show a more pronounced skew toward small p-values. This indicates that the normality tests are more frequently rejecting the null hypothesis, which is expected as a larger portion of the sample no longer follows a normal distribution. Compared to the 10% contamination case, the departure from uniformity becomes more evident and consistent across the tests. Besides, it can be seen that both histograms have a similar behavior regardless the sample size. Therefore, it can be useful to only consider 10^3 iterations due to the computational cost.



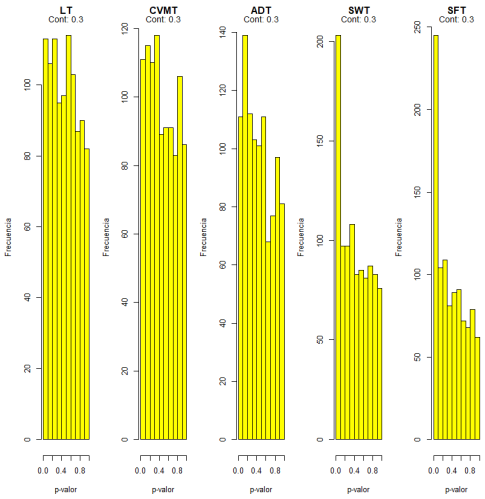
(a) Sample size (n) = 240. Iterations = 1,000



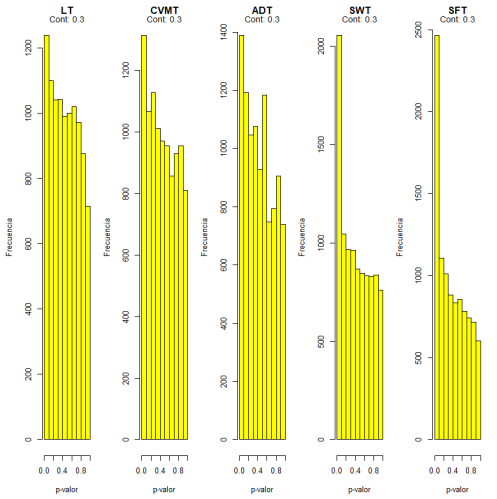
(b) Sample size (n) = 240. Iterations = 10,000

Figure 4: Histograms of p-values under 20% contamination level.

Figure 5 shows the results when the contamination level reaches 30%. In this case, a substantial portion of the data comes from a non-normal distribution. As reflected in the left (5a) and right (5b) panels, the p-value distributions are heavily concentrated near zero. This strong skew confirms that the tests are now consistently detecting non-normality, as the assumption under H_0 is clearly violated. The histograms indicate high power of the tests under this level of contamination. Mainly SWT and SFT are more sensitive to the non-normal evidence, this is shown in the large frecuency of small p-values.



(a) Sample size (n) = 240. Iterations = 1,000



(b) Sample size (n) = 240. Iterations = 10,000

Figure 5: Histograms of p-values under 30% contamination level.

As reflected in the histograms above, all tests increase their rejection rates as the level of contamination and sample size increase. This confirms that the tests are sensitive to deviations from normality. The Shapiro–France tests (SFT) and Shapiro–Wilk tests (SWT) demonstrate greater power, detecting deviations from normality even at low levels of contamination. In contrast, the LT and CVMT tests are more conservative and less sensitive to small deviations, which can be useful for avoiding false positives, but at the cost of lower power.

4 Conclusion

The simulation results underscore the varying sensitivities of normality tests under different levels of data contamination and sample sizes. Overall, the Shapiro–Wilk (SWT) and Shapiro–Francia (SFT) tests displayed the highest power in identifying non-normality as contamination increased, particularly with larger sample sizes. For example, with 30% contamination and $n = 960$, the rejection rates for these tests were approximately 36.89% and 43.37% respectively, indicating a substantial capacity to detect deviations from the null hypothesis. In contrast, tests like Lilliefors (LT), Cramér–von Mises (CVMT), and Anderson–Darling (ADT) exhibited a more moderate increase in rejection rates, suggesting they are less sensitive to mild contamination but still useful when sample sizes are large.

This behavior is consistent with statistical theory, since the null hypothesis H_0 assumes that the data come from a normal distribution:

$$H_0 : X \sim N(\mu, \sigma^2),$$

whereas the alternative hypothesis H_1 posits that the data do not follow a normal distribution:

$$H_A : X \not\sim N(\mu, \sigma^2).$$

Under increasing contamination from a t-distribution, the true distribution of the data deviates from normality, making H_0 less plausible. However, if the data set is small or only slightly contaminated, more conservative tests like LT or CVMT may help prevent the rejection of the null hypothesis.