Okay, so I need to address the problem of inflated Type I error rates when performing a t-test only on data that passed the Shapiro-Wilk normality test. This "double dipping" or selective inference issue is something I've heard about before, but I need to really understand it and find a remedy based on the article mentioned. Let me break this down step by step.

First, I should recap the problem. When researchers first test their data for normality using something like the Shapiro-Wilk test and then proceed with a t-test only if the data passes the normality test, they're introducing a selection bias. This two-step process affects the Type I error rate because the decision to use the t-test is conditional on the Shapiro-Wilk result. Essentially, by filtering out non-normal data, the remaining "normal" datasets might not be truly normal in a way that satisfies the t-test's assumptions, leading to incorrect Type I error rates.

The literature refers to this as "selective inference," where the statistical test is applied after a selection process, which can invalidate the test's assumptions. The challenge here is to adjust the statistical procedure to account for this selection step, thereby controlling the Type I error rate appropriately.

Now, the user mentioned an article that might present approaches to remedy this. Since I can't access the article, I'll have to rely on known methods in the statistical literature that address selective inference or pre-testing bias. Common approaches include:

1. \*\*Adjusting the significance level\*\*: Modifying the alpha level to account for the pre-test, though this can be tricky and may not fully solve the problem.

2. \*\*Using robust tests\*\*: Employing statistical tests that are less sensitive to normality assumptions, such as non-parametric tests (e.g., Wilcoxon rank-sum test), which don't require normality.

3. \*\*Bootstrap methods\*\*: Using resampling techniques to estimate the distribution of the test statistic under the null hypothesis, which can be more reliable when data has been selected based on a pre-test.

4. \*\*Conditional inference\*\*: Adjusting the inference to account for the selection process. This might involve calculating p-values or confidence intervals conditional on passing the pre-test.

5. \*\*Combining tests\*\*: Integrating the pre-test and main test into a single procedure that adjusts for the selection, possibly using a combined hypothesis testing framework.

6. \*\*Bayesian approaches\*\*: Incorporating prior knowledge or using Bayesian methods that can naturally account for model selection uncertainty.

Given that the problem here is specifically about pre-testing for normality before applying a t-test, I should focus on methods that adjust the t-test's procedure to account for the selection bias introduced by the Shapiro-Wilk pre-test.

One approach I recall is the "Two-Stage" procedure, where the first stage is the normality test, and the second stage is the t-test or a non-parametric test. However, this approach doesn't adjust for the selection bias. Instead, methods like the "Conditional t-test" adjust the critical values or p-values of the t-test based on the outcome of the pre-test.

Another method is using a weighted combination of the t-test and a non-parametric test, where the weights depend on the result of the pre-test. This might help balance the Type I error control.

Alternatively, there's the "Adjusted Alpha" approach, where the significance level for the t-test is adjusted downward to compensate for the pre-test. For example, if the pre-test is conducted at α1 and the main test at α2, the overall Type I error rate might be controlled at α = α1 + α2. However, this is simplistic and may not be accurate.

A more sophisticated method is the "Conditional Error Rate" approach, which calculates the actual Type I error rate given that the data passed the pre-test. This involves integrating over the distribution of the test statistics under the null hypothesis, conditional on passing the pre-test. This can be complex but provides a more accurate adjustment.

Bootstrap methods could also be useful here. By resampling from the data and performing the pre-test and t-test on each resample, one can estimate the true distribution of the t-test statistic under the null hypothesis, accounting for the selection process. This empirical distribution can then be used to calculate adjusted p-values.

Additionally, there's the "Selective Inference" framework, which explicitly models the selection process and adjusts the inference accordingly. This area has been developing recently, particularly in high-dimensional statistics and post-selection inference.

Given that the user mentioned an article, perhaps one of these methods is discussed there. Since I can't access the article, I'll outline a plausible remedy based on existing literature.

A practical approach might involve using a permutation test or a bootstrap method that adjusts for the pre-test. Here's a rough outline:

1. \*\*Perform the Shapiro-Wilk test\*\* on the data.

2. \*\*If the data passes\*\* the normality test (p > α\_pre), proceed with the standard t-test.

3. \*\*If the data fails\*\* the normality test, use a non-parametric test like the Wilcoxon test.

4. \*\*Adjust the p-value\*\* from the t-test or Wilcoxon test to account for the pre-test selection. This adjustment can be done using a bootstrap or permutation method that simulates the joint distribution of the pre-test and main test under the null hypothesis.

Alternatively, using a robust test like the Welch t-test, which is less sensitive to normality, might mitigate the issue without needing a pre-test. However, if the pre-test is a necessary part of the analysis protocol, then adjusting for it is essential.

Another angle is to use a Bayesian t-test that incorporates the uncertainty from the pre-test into the posterior probabilities. This would involve specifying prior probabilities for the data being normal or not and updating them based on the Shapiro-Wilk result.

In terms of implementation, a bootstrap-based adjustment might be feasible. Here's how it could work:

- \*\*Step 1\*\*: Compute the Shapiro-Wilk p-value on the original data.

- \*\*Step 2\*\*: Decide to use t-test (if p > α\_pre) or another test (if p ≤ α\_pre).

- \*\*Step 3\*\*: Compute the original test statistic (t-test or alternative).

- \*\*Step 4\*\*: Generate bootstrap samples under the null hypothesis (e.g., by centering the data to have the same mean).

- \*\*Step 5\*\*: For each bootstrap sample:

- Perform the Shapiro-Wilk test.

- Apply the same decision rule as in Step 2.

- Compute the test statistic for the chosen test.

- \*\*Step 6\*\*: Construct the null distribution of the test statistic, accounting for the selection process.

- \*\*Step 7\*\*: Compare the original test statistic to this adjusted null distribution to compute the corrected p-value.

This method would account for the variability introduced by the pre-test and provide a p-value that controls the Type I error rate appropriately.

Alternatively, a simpler approach might be to use a unified test that doesn't require pre-testing. For example, the Brunner-Munzel test is a non-parametric test that can handle heteroscedasticity and non-normal data without needing a pre-test.

However, if the pre-test is a necessary step (e.g., due to regulatory requirements), then adjusting the inference post-selection is necessary. The key is to model the dependency between the pre-test and the main test and adjust the p-value accordingly.

In conclusion, the remedy would involve either:

1. \*\*Adjusting the inference\*\* (p-values or critical values) to account for the selection process using resampling methods or conditional probability calculations.

2. \*\*Using a robust test\*\* that doesn't rely on normality assumptions, thereby eliminating the need for pre-testing.

3. \*\*Bayesian methods\*\* that incorporate model uncertainty from the pre-test into the final inference.

The exact implementation would depend on the specific context and the availability of computational resources for methods like bootstrapping. The adjusted approach should be validated through simulation studies to ensure that it effectively controls the Type I error rate under various data distributions.