The Journal of Open Source Software

Statmanager-kr: A user-friendly statistical package for python in pandas

Changseok Lee [1]

1 DYPHI Research Institute, DYPHI Inc., Daejeon, Korea

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 15 January 2024
Published: unpublished

Summary

Python is one of the most popular and easiest programming languages to learn and use. However, despite many people using Python for statistical analysis, it is difficult to find a statistical package that can match Python's user-friendly nature within the context of statistical analysis. Consequently, people who possess statistical knowledge but lack familiarity with programming languages continue to rely on other costly and inconvenient software. The `statmanager-kr` was designed to provide easy-to-use statistical functions, even for people with little knowledge of programming languages. Because many people are already familiar with data in table format, such as that in Microsoft Excel, `statmanager-kr` was designed to be compatible with `Pandas.DataFrame`. Additionally, statmanager-kr relies on `scipy` and `statsmodels` for accurate and valid statistical analysis. The `statmanager-kr` provides functions related to testing for normality and homoscedasticity assumptions, comparing between-group and within-group differences, conducting regression analysis, and data visualization.

Statement of need

The `statmanager-kr` is an statistical package for Python in `Pandas`. This package provides functions that have commonly used for null hypothesis significance testing (NHST), which is of interest to researchers in various research fields (Moon, 2020). The `statmanager-kr` provides statistical analysis functions to test for significant differences between groups in specific data or within groups in data collected multiple times, based on the researcher's or student's hypothesis. Before applying a specific analysis, it is possible to check whether the normality assumption or the equivariance assumption is met, depending on the distribution of the data. For example, the Shapiro-Wilk test can be used to assess the assumption of normality. To verify the assumption of equality of variances, either the Levene test or the Fmax test can be utilized. Based on these results, the hypothesis of interest can be tested by conducting an independent samples T-test or Mann-Whitney U test. This function helps researchers determine whether the hypothesis formulated by the researcher is acceptable or not.

Most statistical software available to date is difficult to use, inconvenient, and comes with a high cost. In fact, a previous study reported that one of the difficulties university students face in methodological courses like statistics was the "hands-on" exercises, which involve using software (Murtonen & Lehtinen, 2003). Although the inherent difficulty of statistics may be unavoidable, the low usability of the software can be addressed. To achieve this goal, the `statmanager-kr` was designed to enable the application of all analysis with just three lines of codes: 1. reading data as a `Pandas.DataFrame`, 2. creating a `Stat_Manager` object, 3. running `.progress()` method. Therefore, users can use the `statmanager-kr` as long as they learn a minimum of `Pandas` methods to read the data, such as `.read_csv()`, or `.read_excel()`. In addition, it includes additional functions like `.figure()` to visualize results in commonly used

ways depending on the analysis method.

Features

The `statmanager-kr` was designed to be compatible with the wide range form of `pandas.DataFrame`. The implementation of analysis methods and purposes in statmanager-en can be summarized as follows.

Objective	Analysis
Check the normality assumption	Kolmogorov-Smirnov Test, Shapiro-Wilks Test, Z-Skeweness & Z-Kurtosis Test
Check the homoskedasticity assumption	Levenve Test, Fmax Test
Frequency analysis	Chi-Squared Test, Fisher's Exact Test
Check the reliability of the scale	Calculating Cronbach's Alpha
Correlation analysis	Pearson's r, Spearman's rho, Kendall's tau
Comparison between groups	Independent Samples T-test, Yuen's T-test, Welch's T-test, Mann-whitney U test, Brunner-Munzel Test, One-way ANOVA, Kruskal Wallis Test, One-way ANCOVA
Comparison within group	Dependent Samples T-test, Wilcoxon-Signed Rank Test, One-way Repeated Measures ANOVA, Friedman Test, Repeated Measures ANCOVA,
Comparison by multiple ways	N-way ANOVA, N-way Mixed Repeated Measures ANOVA
Regression analysis	Linear Regression, Logistic Regression
etc	Bootstrapping percentile method

All analysis method has its own "key" that enables its application in the `.progress()` method. The analysis is conducted by providing the "key" for each analysis method to the `method` parameter in the `.progress()`, the variables to be analyzed to `vars` parameter, and the group variables to `group_vars` parameter.

```
import pandas as pd
from stamanager import Stat_Manager

df = pd.read_csv(r'../testdata.csv', index_col = 'name')              # 1. Reading the
sm = Stat_Manager(df)                                                 # 2. Creating obj
sm.progress(method = 'ttest_ind', vars = 'weight', group_vars = 'sex') # 3. Running: che
```

Also, if a post-hoc test is required, as in the case of a one-way ANOVA (key of one-way ANOVA is `f_oneway`), it can be conducted by simply providing `True` to the `posthoc` parameter.

```
#Omit the import syntax
df = pd.read_csv(r'../testdata.csv', index_col = 'name')
sm = Stat_Manager(df)

# check the differences in income by condition
sm.progress(method = 'f_oneway', vars = 'income', group_vars = 'condition', posthoc = Tr
```

Keys and Related Informations

The method-specific information required to use the `.progress()` method can be found by using the `.howtouse()` method. The detailed information is summarized in the table below:

Key	Analysis	Required Parameters	Optional Parameters
`kstest`	Kolmogorov-Smirnov Test	`vars`	`group_vars`
`shapiro`	Shapiro-Wilks Test	`vars`	`group_vars`
`z_normal`	Z-skeweness & z-kurtosis test	`vars`	`group_vars`
`levene`	Levene Test	`vars, group_vars`	
`fmax`	Fmax Test	`vars, group_vars`	
`chi2_contingency`	Chi-squared Test	`vars`	
`fisher`	Fisher's Exact Test	`vars`	
`pearsonr`	Pearson's r	`vars`	
`spearmanr`	Spearman's rho	`vars`	
`kendallt`	Kendall's tau	`vars`	
`ttest_ind`	Independent Samples T-test	`vars, group_vars`	
`ttest_rel`	Dependent Samples T-test	`vars`	
`ttest_ind_trim`	Yuen's Two Samples T-test	`vars, group_vars`	
`ttest_ind_welch`	Welch's Two Samples T-test	`vars, group_vars`	
`mannwhitneyu`	Mann-Whitney U Test	`vars, group_vars`	
`brunner`	Brunner-Munzel Test	`vars, group_vars`	
`wilcoxon`	Wilcoxon-Signed Rank Test	`vars`	
`bootstrap`	Boostrap Percentile Method	`vars`	`group_vars`
`f_oneway`	One-way ANOVA	`vars, group_vars`	`posthoc, posthoc_method`
`f_oneway_rm`	One-way Repeated Measures ANOVA	`vars`	`posthoc, posthoc_method`
`kruskal`	Kruskal-Wallis Test	`vars, group_vars`	`posthoc, posthoc_method`
`friedman`	Friedman Test	`vars`	`posthoc, posthoc_method`
`f_nway`	N-way ANOVA	`vars, group_vars`	`posthoc, posthoc_method`
`f_nway_rm`	N-way Mixed Repeated Measures ANOVA	`vars, group_vars`	`posthoc, posthoc_method`
`linearr`	Linear Regression	`vars`	
`hier_linearr`	Hierarchical Linear Regression	`vars`	
`logisticr`	Logistic Regression	`vars`	
`oneway_ancova`	One-way ANCOVA	`vars, group_vars`	
`rm_ancova`	One-way Repeated Measures ANCOVA	`vars`	
`cronbach`	Calculating Cronbach's Alpha	`vars`	

Also, the `statmanager-kr` provide two posthoc methods. It can be conducted by providing key of `posthoc_method` parameter as belows:

Key of `posthoc_method`	Method
`bonf`	Bonferroni Correction
`tukey`	Tukey HSD

Acknowledgements

Author declares no conflicts of interests.

References

Moon, S. M. (2020). *Statistics for the social sciences: Moving toward an integrated approach*. Cognella Academic Publishing. ISBN: 978-1516519613

Murtonen, M., & Lehtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Studies in Higher Education*, *28*(2), 171–185. https://doi.org/10.1080/0307507032000058064