User's Manual for NLPStatTest

September 6, 2020

1 Introduction

This is the user's manual for the system NLPStatTest, where we will provide instructions on how to run this system and definitions of parameters/arguments the users need to specify.

1.1 How to run NLPStatTest?

There are three options to run NLPStatTest:

Online

NLPStatTest can be run from the website https://nlpstats.ling.washington.edu/, if the users have a reliable internet connection.

Local GUI

The users can choose to download the system and install it on their own computers, which probably will require installation of additional Python packages.

Command Line

The users can also choose to directly run the system using command line by calling Python.

2 Definitions

In this section, we provide basic definitions for statistical terms that are used in this system. NLPStatTest is based on the frequentist approach to hypothesis testing, often known as the null hypothesis testing framework.

Definition 1 (**Null hypothesis).** The **null hypothesis** of statistical hypothesis testing, usually denoted by H_0 (pronounced as *H-naught*), is the default hypothesis which usually connotes to the absence of a phenomenon of interest.

Definition 2 (**Alternative hypothesis).** The **alternative hypothesis**, denoted by H_1, is the complementary hypothesis for the null hypothesis, which usually means the presence of a phenomenon.

H_0 and H_1 are usually mutually exclusive and phrased with respect to some fixed parameters on the population level such as the mean or median of some probability distribution. For example, we may want to investigate whether two samples have the same mean or not. The corresponding hypotheses are:

$$H_0 : \ \mu_X = \mu_Y \ \ v.s. \ \ H_1 : \ \mu_X \neq \mu_Y \tag{1}$$

where μ_X and μ_Y denote the means of samples X and Y respectively.

We will employ a statistical significance test to test this hypothesis, which will produce a p-value.

Definition 3 (P**-value).** The p**-value** of a significance test is the probability that under H_0 the test statistic is at least as extreme as the observed one.

$$p = P\big(T \geq t | H_0\big) \tag{2}$$

where T is the test statistic and t is the observed value of T.

Definition 4 (**Test statistic).** The **test statistic** T of a significance test is a function of the sample that is used to determine whether the null hypothesis should be rejected.

Since T is a function of the sample, it is a random variable and thus follows a probability distribution, also known as the sampling distribution. The distribution of T when H_0 is true is called the null distribution. The observed value of T tends to be too large or too small with respect to the null distribution if the null hypothesis is false.

When a p-value is given, we then can draw a conclusion with respect to the null hypothesis. Before conducting a significance test, a significance level should be specified, which is known as the Type I error of a test.

Definition 5 (**Type I error).** The **Type I error** of a significance test is the probability that under H_0, H_0 is rejected by the test, which is usually denoted by α.

$$\alpha = P\big(H_0 \; is \; rejected \mid H_0 \; is \; true\big) \tag{3}$$

Definition 6 (**Type II error).** The **Type II error** of a significance test is the probability that under H_1 (the alternative hypothesis), the alternative is rejected by the test, which is usually denoted by β.

$$\beta = P\big(H_1 \; is \; rejected \mid H_1 \; is \; true\big) \tag{4}$$

Type I and II errors are the two errors we wish to control when running a statistical test. Type I error can be easily controlled by presetting α, while Type II error can be controlled by conducting power analysis.

3 Interpretations

This section deals with proper interpretations of testing results such as the p-value and confidence intervals.

3.1 P-value

The p-value measures how incompatible the current data is with a proposed statistical model or hypothesis. It does not indicate whether H_0 is false or true, nor does it measure the importance or size of an effect. Given a significance level α, we can draw a conclusion as follows:

1. If p-value $< \alpha$, we reject H_0.
2. If p-value $\geq \alpha$, we **fail to reject** H_0.

Note that we can never **accept** a hypothesis.

3.2 Confidence interval

A confidence interval is an interval estimation of the parameter of interest (e.g. mean, median). It accompanies the point estimation of the parameter of interest and provides a measure of uncertainty. The proper interpretation of a 95% confidence interval is as follows:

> If we repeatedly resample from the population and calculate the confidence interval for many times, then 95% of the intervals will contain the true value of the parameter.

Note that it is a common misconception that the confidence interval is the interval that will contain the true value of the parameter with 95% probability.

4 Procedures

The online version of the system `NLPStatTest` contains four steps: **Data Analysis**, **Significance Testing**, **Effect Size** and **Power Analysis**, with an optional step for pre-testing power analysis.

4.1 Pre-testing Power Analysis

4.2 Data Analysis

The first step is exploratory data analysis, where the users need to upload the data file. The system will compute summary statistics (mean, median, standard deviation etc), plot histograms, run normality test, skewness check and finally recommend a list of appropriate significance tests.

Data File

In the first step, the users need to upload the data file they wish to analyze. The data file should only contain two columns of numbers, separated by whitespace. For example,

1.1373 -0.4661
0.7997 1.4805
0.3074 -0.0963
1.6159 -0.2737
1.5926 -0.5972
. . .

Evaluation Unit Size

After uploading the data file, the users need to specify the **evaluation unit size** (EU size). The **evaluation unit** is a set of data points on which the chosen evaluation unit is meaningfully defined.

For example, in ML evaluation, usually the sentence-level BLEU scores do not provide a reliable measure for translation quality, while an average of multiple sentence-level BLEU scores or a corpus-level BLEU score can better approximate translation quality. In this case, if the uploaded data file contains sentence-level BLEU scores and the users decide that the average of 15 scores is a reliable measure, then the EU size is 15.

Evaluation Unit Metric

Then, the users need to choose how they want to calculate one evaluation unit, either by mean or median. This is called the **evaluation unit metric**.

Random Seed and Reshuffling

5 Technical details

In this section, we will present technical details of significance tests, effect size indices and power analysis.

5.1 Significance tests

5.1.1 Shapiro-Wilk test

The Shapiro-Wilk normality test is a significance test for testing normality of the data.

5.1.2 Skewness test*

The skewness test is not a significance test but a rule-of-thumb check for whether the distribution of the data is skewed.

5.1.3 Student t test

The student t test is a classic significance test for comparing means.

5.1.4 Sign test

The (exact) sign test

5.1.5 Wilcoxon signed-rank test

5.1.6 Bootstrap test

5.1.7 Permutation test

5.2 Effect size indices

5.2.1 Cohen's d and Hedges' g

5.2.2 Wilcoxon r

5.2.3 Hodges-Lehmann estimatior

5.3 Power analysis