# User's Manual for NLPStatTest

November 24, 2020

## Contents

# 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 What Is `NLPStatTest` for?

System performance evaluation constitutes a vital part in the NLP field, and it is common practice to resort to statistical significance testing in order to demonstrate that system difference exhibited by the proposed system from the baseline is not due to mere happenstance.

While most of previous work covers significance testing for NLP system comparison, it is rarely pointed out that the $p$-value alone does not suffice for statistical hypothesis testing: For a large enough dataset, the $p$-value will always approach 0 given an extremely tiny but nonzero difference, yielding statistical significance, while the difference might not be scientifically meaningful. We need to consider practical significance as well, which can be measured by estimating effect size. In addition, to ensure the statistical test has enough power (minimizing Type II error), we also need to conduct power analysis when choosing the size of test corpus.

`NLPStatTest` is a statistical tool developed for comparing NLP system performances using statistical significance testing.

We propose a three-stage procedure for comparing NLP system performance. The first stage is building an NLP system. The second stage is hypothesis testing. The last stage is to report various results produced by the second stage. This toolkit is developed for the second stage.

## 1.2 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.

## 1.3 How to Cite `NLPStatTest`?

*@inproceedings{zhu-etal-2020-AACL,*
*title = "NLPStatTest: A Toolkit for Comparing NLP System Performance",*
*author = "Haotian Zhu and Denise Mak and Jesse Gioannini and Fei Xia",*
*booktitle = "Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)",*
*publisher = "Association for Computational Linguistics",*
*year = 2020,*
*month = "December"*
}

## 1.4 Contact Information

If the users encounter any technical difficulties or have suggestions for improving `NLPStatTest` please email all of the authors at:

{haz060, dpm3, jessegio, fxia}@uw.edu

# 2 Prospective Power Analysis

In `NLPStatTest`, prospective power analysis (Figure 1) is a preliminary and optional step to calculate minimal sample size for a normal sample to achieve a certain level of statistical power. The user needs to provide the expected mean and standard deviation of the differences between samples, the desired power level, and the required significance level. `NLPStatTest` will calculate the minimally required sample size for $t$ test via a closed form, **assuming the normal distribution of the data**.

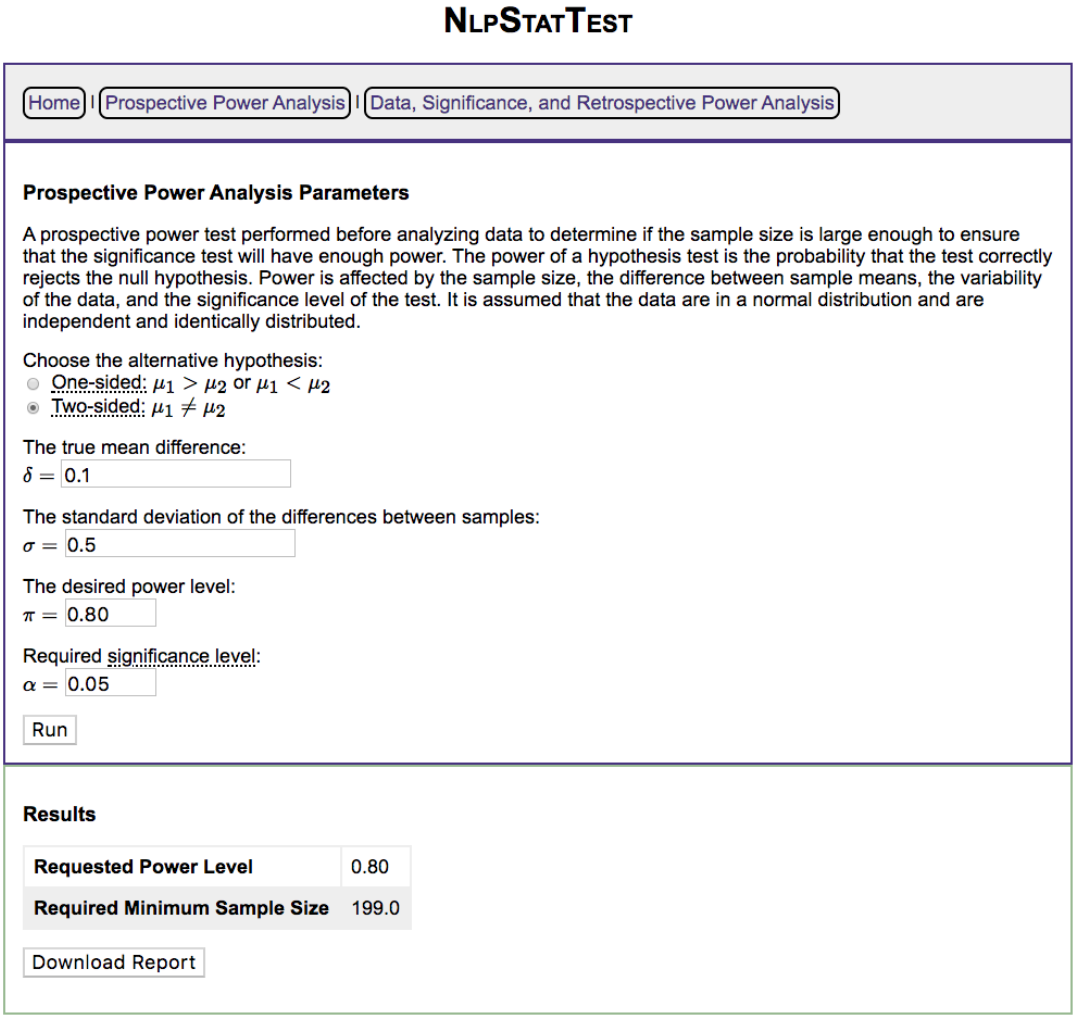



Figure 1: Prospective power analysis

## 2.1 System Specification

- ***Sidedness***: whether the alternative hypothesis is one-sided or two-sided.
- $\delta$: true difference of test statistic between the two samples.
- $\sigma$: standard deviation of the differences between the two samples.
- $\pi$: desired power level.
- $\alpha$: significance level.

## 2.2 System Output

After specifying those parameters and hitting the *run* button, the system will output an integer number indicating that a sample size equal or larger than the number will achieve the desired power level (Figure 1).

# 3 Data, Significance and Retrospective Power Analysis

The online version of the system `NLPStatTest` contains four primary steps: **Data Analysis**, **Significance Testing**, **Effect Size** and **Retrospective Power Analysis**.

The preliminary step to the main comparison procedure is to upload the data file (Figure 2) and the config file (optional). 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
. . .

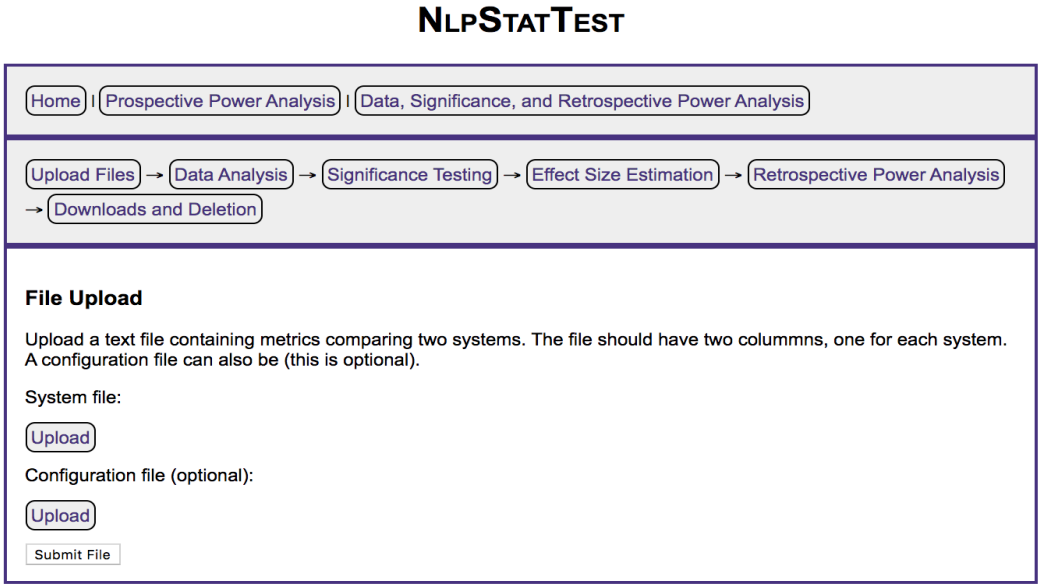



Figure 2: File uploading page

## 3.1 Data Analysis

The first step is exploratory data analysis (Figure 3), where the users need to upload the data file. The system will compute and report summary statistics (mean, median, standard deviation etc), plot histograms for data visualization, run normality test, check for distributional symmetry based on sample skewness and recommend a list of appropriate significance tests.

### 3.1.1 System Specification

- ***Evaluation Unit Size***: the number of data points within an evaluation unit. Must be a positive integer.
- ***Evaluation Unit Metric***: the metric to compute an evaluation unit. Can be either mean or median.
- ***Random Seed***: the random seed to reshuffle the data before converting the data to evaluation units.
- ***Significance Level $\alpha$ (for calculating normality)***: the significance level for running the Shapiro-Wilk normality test.

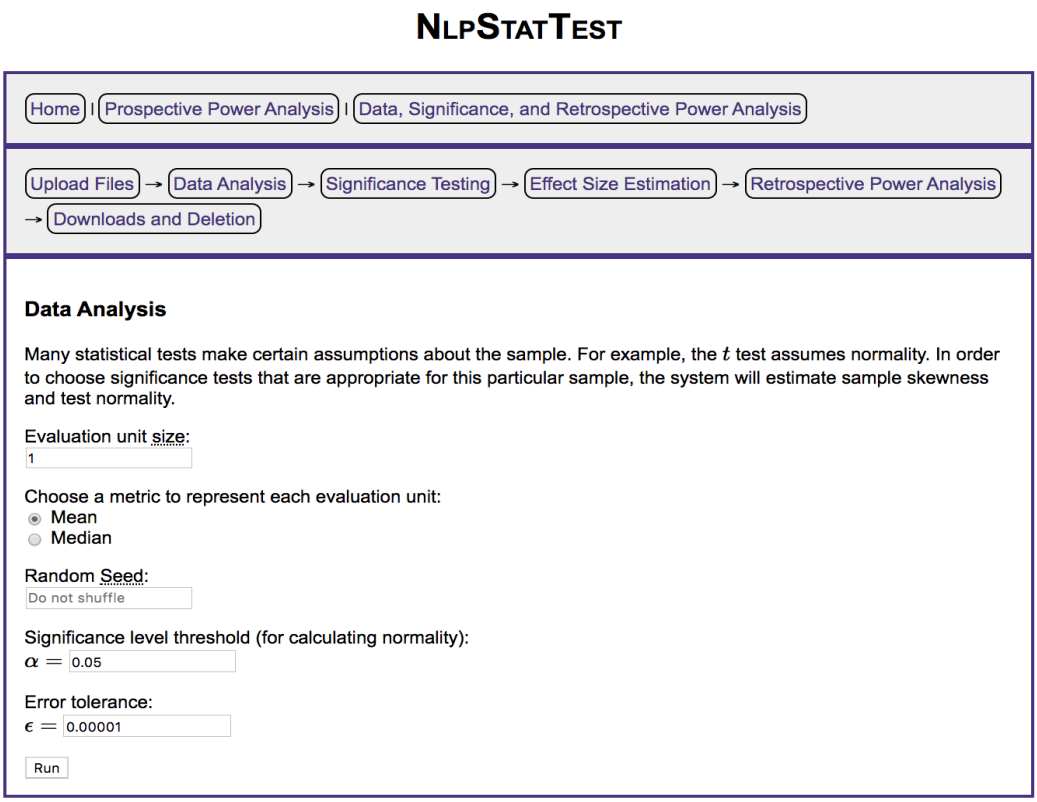



Figure 3: Data analysis page

### 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**. Note that the evaluation unit metric is different from the evaluation metric used to quantify system performance difference.

### Random Seed and Reshuffling

Sometimes the users might want to reshuffle the data first before testing, if they wish to get rid of unnecessary sequential dependency that could potentially exist within the dataset. They just need to specify the random seed used for reshuffling. The default choice is to not reshuffle.

### Normality Test

The last step is to specify the significance level $\alpha$ for normality test which will check if the data's distribution is statistically significantly different from a normal distribution. There are many normality tests: Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, Pearson, etc. The test used in this step is the Shapiro-Wilk normality test.

Note that the Python built-in function which implements the Shapiro-Wilk test will produce a warning that the $p$-value might not be accurate if the sample size is larger than 5000.

### 3.1.2 System Output

After uploading the data file and specifying the aforementioned system parameters, hit the *Run* button and the results for data analysis will be shown (Figure 4).

#### Summary of Statistics

First, summary statistics for the evaluation units will be shown. *score1* and *score2* are the EUs for the first and second columns respectively, and *difference* is computed by *score1* minus *score2*.

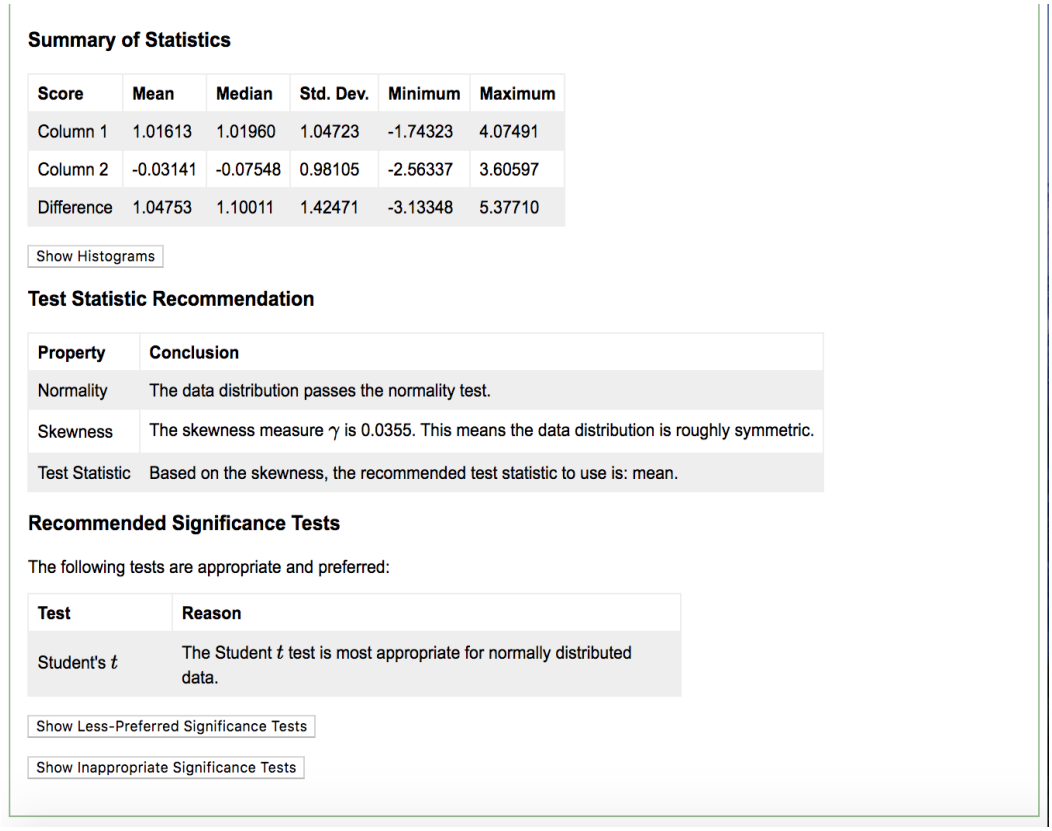



Figure 4: Data analysis output

#### Test Statistic

By conducting normality test and checking for skewness, `NLPStatTest` will choose a test statistic, mean or median, to measure the central tendency.

#### Histograms

Under the summary statistics table, there is a button *View Histograms*, which the users can hit to show the histograms of the three scores.

#### Recommended Tests

`NLPStatTest` will return three lists of significance tests: recommended tests, less-preferred tests and inappropriate tests. The algorithm for choosing a test is given in Appendix **??**. Within the list, each test is followed by a case-by-case reason for why this test is recommended, less-preferred or inappropriate given the input data.

## 3.2 Significance Testing

The second step is significance testing (Figure 5), where users need to choose a significance test from the lists of recommended tests or less-preferred tests. The difference between these two lists is that the less-preferred tests satisfy the assumptions but may lack computational efficiency (e.g. tests based on randomization) or have less statistical power (e.g. nonparametric, rank-based tests). By specifying the required parameters, this step will output the name of the test, a confidence interval for the chosen test statistic, a $p$-value and the decision to reject or fail to reject $H_0$.

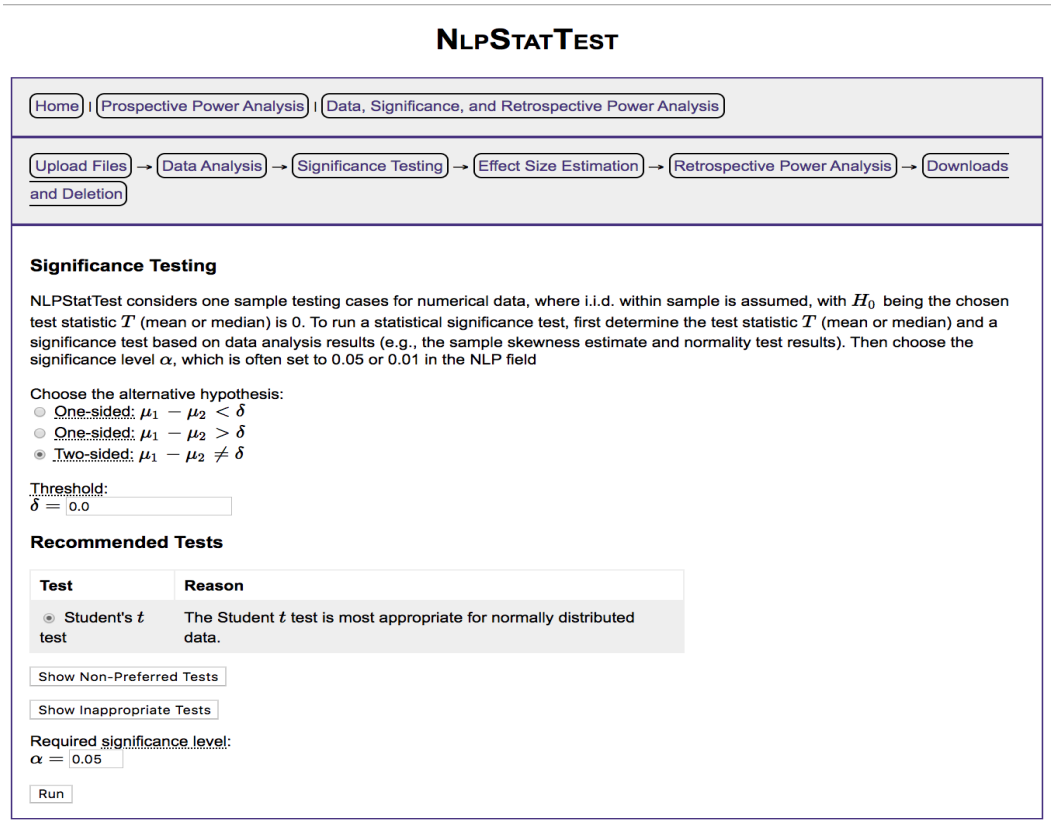



Figure 5: Significance testing page

### 3.2.1 System Specification

- ***Alternative Hypothesis***: the directionality of the alternative hypothesis: less, greater or two-sided.
- $\delta$: the hypothesized difference of the test statistic
- ***Significance Test***: a significance test from the list of recommended or less-preferred.
- ***Significance Level*** $\alpha$: the significance level for running significance test.
- ***Number of Iterations (optional)***: the number of iterations for bootstrap or permutation significance tests.

**Directionality of** $H_1$

There are three options to specify $H_1$: **less**, **greater** or **two-sided**:

1. $\mu_1 < \mu_2$
2. $\mu_1 > \mu_2$
3. $\mu_1 \neq \mu_2$

where $\mu_1$ and $\mu_2$ are the test statistics for the two samples. First two options correspond to one-sided tests; the third option corresponds to a two-sided test. For all three options, $H_0 : \mu_1 = \mu_2$.

**Threshold $\delta$**

$\delta$ is the hypothesized value of differences in the test statistics for the two samples under $H_1$, where there is a difference between $\mu_1$ and $\mu_2$. $\delta$ can be any real numbers.

### 3.2.2 System Output

After specifying required parameters and hitting the *run* button, the result section will show the test name, confidence interval, $p$-value and the decision to reject or fail to reject $H_0$ (Figure 6).

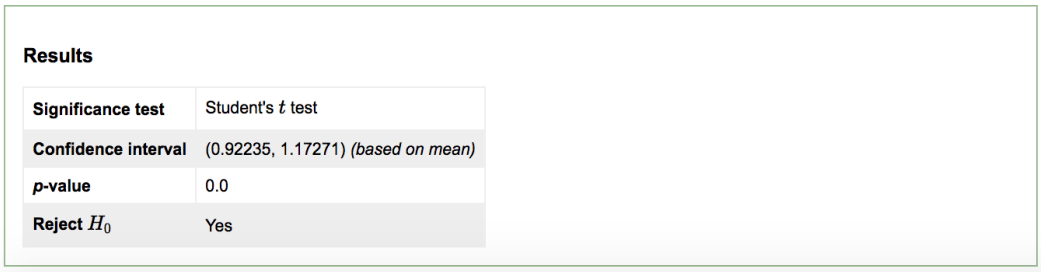



Figure 6: Significance testing output

## 3.3 Effect Size

It is necessary to estimate effect size in order to measure practical significance in addition to statistical significance. `NLPStatTest` provides the estimation of effect size via four choices of effect size indices (Figure 7): the famous Cohen's $d$, Hedges' $g$, Wilcoxon $r$ for the Wilcoxon signed-rank test and the Hodges-Lehmann estimator for median. The first three are standardized estimators, while the last is unstandardized. Additionally, `NLPStatTest` will compute the confidence interval for the chosen estimators. This step is optional in the sense that the users can obtain the comparison result without completing this step, but we strongly advise for estimating effect size for reproducibility of the comparison result and for the sake of meta-analysis.

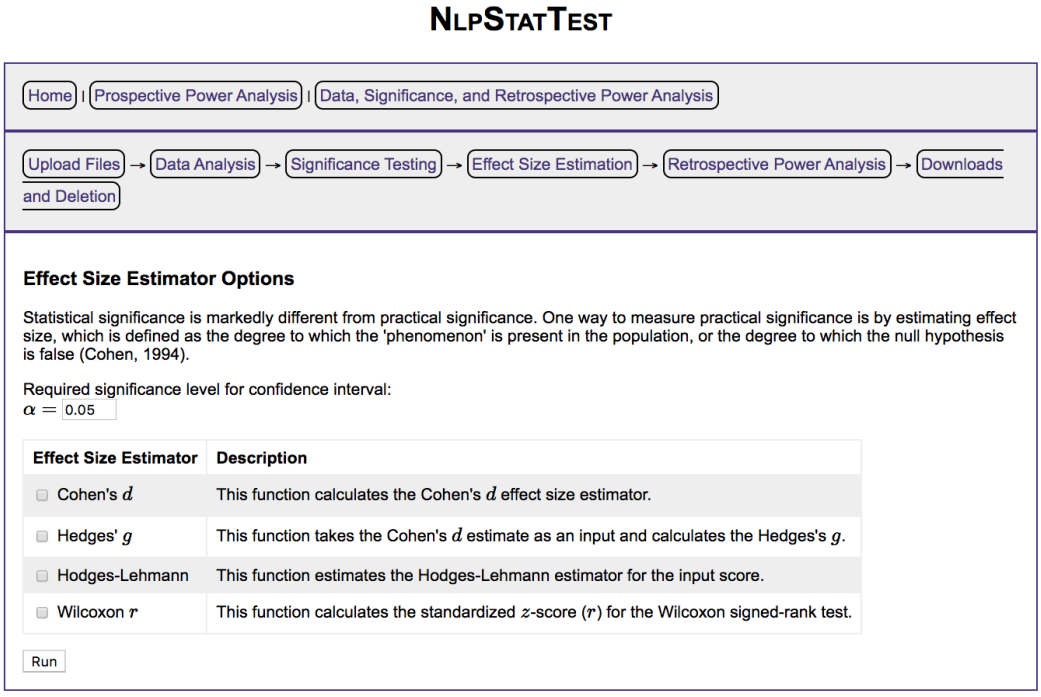



Figure 7: Effect size page

### 3.3.1 System Specification

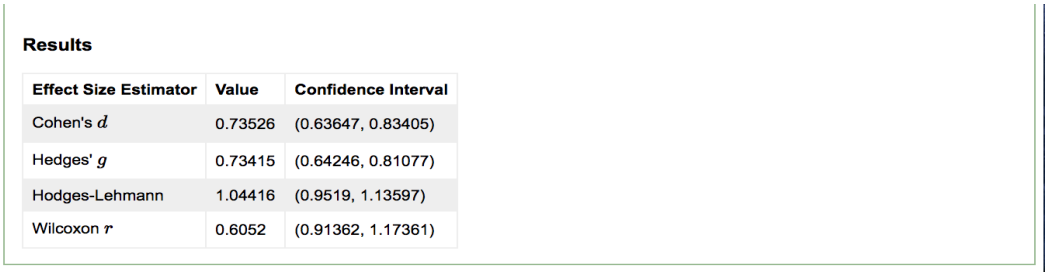
- ***Significance Level $\alpha$***: the significance level for computing the confidence interval for effect size estimates.
- ***Effect Size Estimator***: choose one or more estimators of interest.

#### Effect Size Estimator

The choice of effect size estimator is associated with what significance test the users choose in the previous step. Generally, Cohen's $d$ and Hedges' $g$ are for student $t$ test, Wilcoxon $r$ is for Wilcoxon signed-rank test and Hodges-Lehmann estimator is for any test that is based on median with the assumption that the data distribution is symmetric.

### 3.3.2 System Output

After specifying the required parameters, `NLPStatTest` will display the effect size estimates as well as the confidence intervals (Figure 8).



**Results**

| Effect Size Estimator | Value | Confidence Interval |
|---|---|---|
| Cohen's $d$ | 0.73526 | (0.63647, 0.83405) |
| Hedges' $g$ | 0.73415 | (0.64246, 0.81077) |
| Hodges-Lehmann | 1.04416 | (0.9519, 1.13597) |
| Wilcoxon $r$ | 0.6052 | (0.91362, 1.17361) |

Figure 8: Effect size output

## 3.4 Retrospective Power Analysis

The last step in the main comparison stage is the retrospective power analysis (Figure 9), which is to compute the achieved statistical power for the observed data and to estimate the false negative rate (in fact the achieved power is the true negative rate: how many tests correctly reject $H_0$ if $H_1$ is true). In this step, `NLPStatTest` will ask the users to choose the number of different sample sizes and the number of iterations for each sample size to sample from the distribution in order to perform power analysis. There are two methods for retrospective power analysis: Monte Carlo simulation directly from the known distribution (normal) and the bootstrap, which resamples from the empirical distribution of the given data.

### 3.4.1 System Specification

- ***Number of power measurements***: the number of different sample sizes.
- ***Number of iterations at each sample size***
- ***Method of power analysis***: either Monte Carlo or Bootstrap
- ***MC:*** $\alpha$: the significance level
- ***Bootstrap: significance test***: the significance test used in the previous step/
- ***Bootstrap:*** $\alpha$: the significance level.
- ***Bootstrap: Iterations for bootstrap significance test***: the number of iterations for the bootstrap test (this is different from the bootstrap in the power analysis).

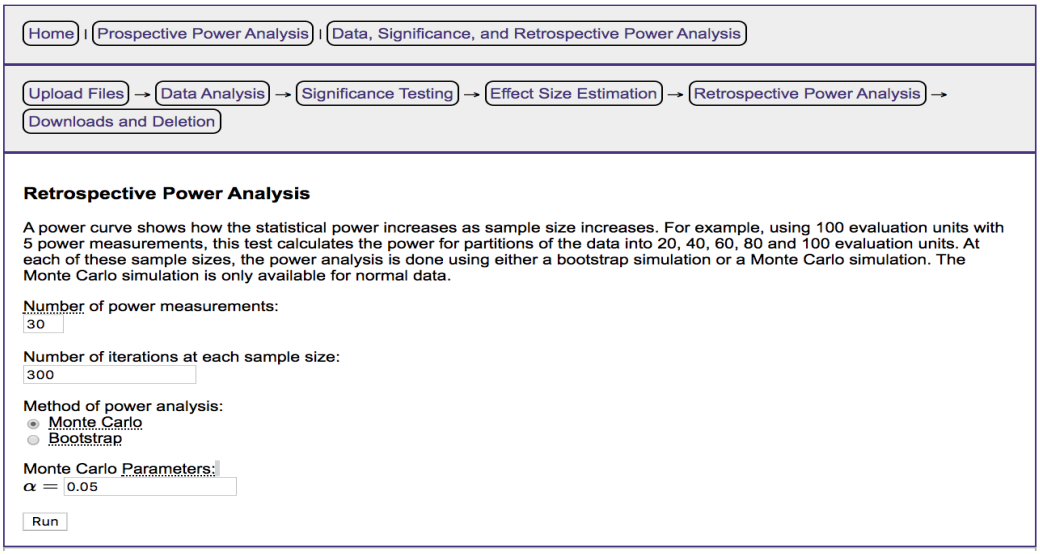



Figure 9: Retrospective power analysis page

### 3.4.2 System Output

After specifying the required parameters, `NLPStatTest` will display a power curve (Figure 10), with different sample sizes on the $x$-axis and statistical power on the $y$-axis.

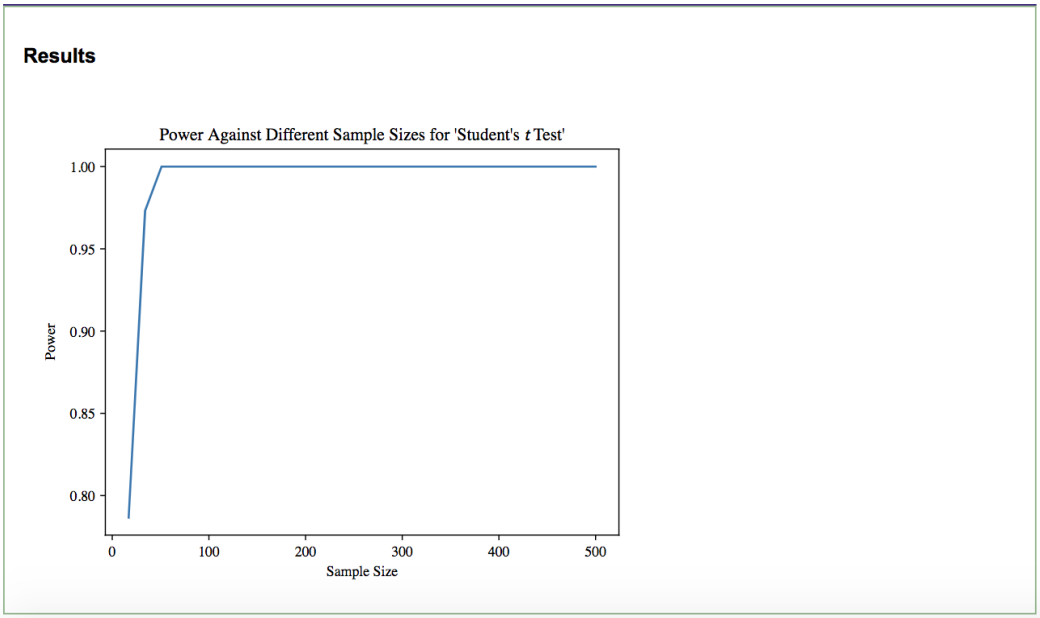



Figure 10: Retrospective power analysis output

# A Appendix: 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.

## A.1 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 $\mu_X$ and $\mu_Y$ 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 $\alpha$.

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

**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 $\beta$.

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

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 $\alpha$, while Type II error can be controlled by conducting power analysis.

**Definition 7** (**Statistical power).** The power of a statistical significance test is the probability that under $H_1$, the null is correctly rejected by the test. The power of a test is $1 - \beta$.

## A.2 Building an NLP System

The first step to compare system performance is to build an NLP system. Here we give definitions of relevant terms used in this paper.

**Definition 8 (Test instance).** A test instance, denoted by $(x, y)$, is a pair, where $x$ is an input to an NLP system and $y$ is the corresponding gold standard for system output.

**Definition 9 (Evaluation unit).** An evaluation unit $e = \{(x_j, y_j), j = 1, \cdots, m\}$ is a set of test instances on which an evaluation metric can be meaningfully defined. A test set is a set of evaluation units.

**Definition 10 (Evaluation metric).** Given an NLP system $A$, the evaluation metric $M$ for the NLP task is a function of an evaluation unit $e$ which produces a numerical value:

$$M_A(e) = M\Big( \{\big(A(x_j), y_j\big),\ j = 1, \cdots, m\} \Big) = M\Big( \{\big(\hat{y}_j, y_j\big),\ j = 1, \cdots, m\} \Big) \tag{5}$$

where $\hat{y}_j = A(x_j)$ is the system output of $A$ given $x_j$ and $m$ is the number of test instances in $e$.

Equation 5 illustrates the evaluation step for an NLP system $A$ by first running the system on $x_j$'s and then comparing $\hat{y}_j$'s with their corresponding gold standard $y_j$. An evaluation unit may contain one or more test instances, and it is essential to understand whether the evaluation metric (Definition 10) is computed using a single test instance (Definition 8) or a set of test instances. For example, a BLEU score can be computed using a set of test sentences or a single sentence. The evaluation metric $M$ maps an evaluation unit to a single numerical value of evaluation metric. Later, we will show that the size of an evaluation unit affects sample size, $p$-value, sample standard deviation, effect size and statistical power.

# B Appendix: Interpretations

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

## B.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 $\alpha$, 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.

## B.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.

| Effect size | $d$ |
|---|---|
| Small | 0.20 |
| Medium | 0.50 |
| Large | 0.80 |

Table 1: Qualitative interpretation of Cohen's $d$

### B.3 Effect Size

There is a set of qualitative interpretations associated with Cohen's $d$ and Hedges' $g$, which is specifically developed for the field of behavioral sciences (by Cohen). The users should consult this with caution since the qualitative interpretation might not be applicable to the NLP field.

### B.4 Power Analysis and Sample Size Computation

The prospective power analysis implemented in `NLPStatTest` assumes the sample is normally distributed and the significance test is the $t$ test. It will compute the required sample size to reach the specified power level. The output is an integer indicating that a sample size at least as large as the output will result in a test which has at least the specified power level.

The retrospective power analysis will produce a power curve with respect to different sample sizes. For a true null, the power is expected to be low for all possible sample sizes; for a true alternative (non-zero effect), the power will increase as the sample size increases. This graph shows that to increase power, one can increase the sample size or increase the effect size.

## C Appendix: Data Analysis

In this appendix, we will explain individual elements of data analysis step in detail.

### C.1 Choosing Evaluation Unit Size

Evaluation unit size is the number of test instances within one evaluation unit, which is the minimal unit from which a value of evaluation metric is computed. The size of evaluation unit matters because it is closely associated with the computation of standard deviation. For small evaluation unit sizes, the sample standard deviation will be large and vary dramatically; it will stabilize after a certain size. In `NLPStatTest`, multiple choices of EU sizes will be used to estimate the sample standard deviation, resulting in a plot showing the relationship between them. It is often the case that after a EU size of 15, standard deviation tends to be stable with indiscernible differences. Users can input a tolerance level to indicate how small the difference will be, and `NLPStatTest` will output the first 10 EU sizes of which the sequential differences are less or equal to that tolerance level.

### C.2 Skewness

The skewness test is not a significance test but a rule-of-thumb check for whether the distribution of the data is skewed. Many statistical tests ($t$ test, bootstrap test based on $t$ ratios, etc) are based on the mean as the test statistic, drawing inferences on average system performance.

However, when the data distribution is not symmetric, the mean does not properly measure the central tendency. In that case, the median is a more robust measure. Another issue associated with mean is that if the distribution is heavy-tailed (e.g., the $t$ and Cauchy distributions), the sample mean oscillates dramatically.

In order to examine the symmetry of the underlying distribution, `NLPStatTest` checks the skewness of $\{u_i - v_i\}$ by estimating the sample skewness ($\gamma$). Based on the $\gamma$ value, we use the following rule of thumb [**?**] to determine whether `NLPStatTest` would recommend the use of mean or median as the test statistic for statistical significance testing:

- $|\gamma| \in [0, 0.5)$: roughly symmetric (use mean)
- $|\gamma| \in [0.5, 1)$: slightly skewed (use median)
- $|\gamma| \in [1, \infty)$: highly skewed (use median)

## C.3 Normality Test

The Shapiro-Wilk normality test is a significance test for testing normality of the data. `NLPStatTest` uses the built-in function of Shapiro-Wilk test in the Python package scipy.stats, where we by default use a two-sided test.

Consider a sample $\mathbf{X} = (X_1, \cdots, X_n)$ from a normal distribution, which is the null hypothesis for the test. The test statistic is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\left(\sum_{i=1}^{n} X_i - \bar{X}\right)^2} \tag{6}$$

where $X_{(i)}$is the $i$-th order statistic (sort $X_i$'s in a ascending order) and $a_i$'s are given by

$$a' = (a_1, \cdots, a_n) = \frac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{1/2}} \tag{7}$$

where the vector $m' = (m_1, \cdots, m_n)$ denotes the vector of expected values of standard normal order statistics and $V$ denotes the corresponding $n \times n$ covariance matrix.

## C.4 Choosing a Significance Test

`NLPStatTest` will provide three lists of significance tests: recommended, non-preferred and inappropriate. The list of recommended tests contains tests that are deemed appropriate and expected to have high statistical power and to have faster computational speed. Non-preferred tests are appropriate tests but may be in lack of statistical power or computationally expensive. Inappropriate tests do not have their assumptions satisfied for the given dataset. Table 2 shows how the three lists are generated depending on the skewness and normality results. Note that permutation and bootstrap here include both mean and median approaches.

| Symmetric | Normal | Test statistic | Recommended | Non-preferred | Inappropriate |
|---|---|---|---|---|---|
| ✓ | ✓ | mean | $t$ | Sign, Wilcoxon, permutation, bootstrap | None |
| ✓ | X | mean | Wilcoxon | Sign, permutation, bootstrap | $t$ (OK for large sample) |
| X | X | median | Sign | permutation (med), bootstrap (med) | $t$, Wilcoxon |

Table 2: Table for choosing a significance test

# D Appendix: Significance Tests

. In this section, we will present details of significance tests implemented in `NLPStatTest`.

| Test name | Symmetric | Normal | Test statistic |
|---|---|---|---|
| $t$ test | ✓ | ✓ | mean |
| Sign test | X | X | median |
| Wilcoxon signed-rank test | ✓ | X | median |
| Bootstrap test | X | X | mean |
| Bootstrap test (median) | X | X | median |
| Permutation test | X | X | mean |
| Permutation test (median) | X | X | median |

Table 3: Table of significance tests implemented by `NLPStatTest` with their assumptions with respective symmetry of distribution and normality and test statistic.

## D.1 Student $t$ Test

The paired student $t$ test is a classic significance test for comparing means. `NLPStatTest` implements $t$ test using a closed form.

The paired student $t$ test is a variant of the regular $t$ test but can be applied to dependent $\mathbf{X}$ and $\mathbf{Y}$, where the sample size should match $n = m$ such that $X_i$ and $Y_i$ form a pair. The assumptions are:

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired and thus dependent.
2. $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed.
3. Each sample consists of $i.i.d.$ observations.
4. Two samples have the same variance.
5. Sample sizes of $\mathbf{X}$ and $\mathbf{Y}$ must be the same.

The testing procedure, in essence, is equivalent to a one-sample student $t$ test by alternatively considering the difference of the original samples, $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$. The test statistic is given by

$$t_{paired} = \frac{\bar{Z}}{S_Z/\sqrt{n}} \tag{8}$$

where $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$ and $S_Z$ denotes the sample standard deviation of $\mathbf{Z}$. Under the null hypothesis, the test statistic $t$ follows a $t$ distribution with $n - 1$ degrees of freedom.

## D.2 Sign Test

The (exact) sign test is a non-parametric test for medians which relies on the binomial distribution. This test does not make any distributional assumption on the data, and thus it has low statistical power due to loss of information which comes from only considering pairwise whether one sample is greater than the other in magnitude. `NLPStatTest` implements the exact sign test based on Nonparametric Statistical Methods (Hollander et al., 2013).

## D.3 Wilcoxon Signed-rank Test

Wilcoxon signed-rank test is a non-parametric test which is usually regarded as the counterpart to the parametric $t$ test. It is testing for medians and assumes symmetry of the distribution of the given sample; otherwise, Wilcoxon signed-rank test does not make any distributional assumption on the data. `NLPStatTest` implements Wilcoxon signed-rank test based on Nonparametric Statistical Methods (Hollander et al., 2013).

Note that the computation of the confidence interval for Wilcoxon signed-rank test tends to be slow for large samples.

## D.4 Bootstrap Test

The bootstrap test is another computational approach which relies on the notion that the given sample approximates the original population and on resampling with replacement. Tests calibrated by the bootstrap can be applied to a even larger variety of hypothesis testing problems than the permutation test. It is basically regarding the given sample as the population and resampling from the given sample with replacement. One of the major concerns in employing a bootstrap method to conduct statistical hypothesis testing is choosing a test statistic. A proper test statistic should exhibit distinguishable differences under the null and alternative hypotheses.

One must note that significance tests based on the bootstrap is computationally expensive when the sample size is extremely large. A simple bootstrap method has approximately $O(n)$ runtime given the sample size is $n$. A bootstrap that has a double loop in order to estimate the standard error of the bootstrapped test statistic has approximately $O(n^2)$ runtime. Also, when the sample size is small (below 100), the sample may not be considered a legitimate surrogate for the entire population, which essentially violates one of the assumptions held by the bootstrap method. In addition, when the sample size is small, the bootstrap cannot be used as a remedy for small sample size.

The assumptions for the bootstrap are:

1. $\mathbf{X}$ and $\mathbf{Y}$ are independent.
2. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are $i.i.d.$.
3. Samples are representative of their populations.

Here, we reduce the paired two-sample testing into a one-sample testing problem by considering the pairwise differences. The test procedure is given as follows:

**Algorithm 1** One-sample Bootstrap Test

**Input:** $X$, the sample
**Output:** $p$, bootstrap $p$-value

**procedure** BOOTSTRAP
  Calculate the observed test statistic of choice $\theta_{ob}(X)$
  Given a large number $B$
  Given a counting number $C \leftarrow 0$
  **for** $b \in 1 : B$ **do**
    Sample $X$ with replacement independently and obtain $X_b$ of the same size as $X$
    Calculate bootstrap test statistic $\theta_b(X_b)$
    **if** $\theta_b \geq \theta_{ob}$ **then**
      $C \leftarrow C + 1$
  Calculate bootstrap $p$ value $= \frac{C+1}{B+1}$
  **return** $p$-value

Note that the $p$-value is given by

$$\frac{C+1}{B+1} \tag{9}$$

where $C$ is the number of extreme bootstrapped test statistics and $B$ is the number of iterations.

As usual, given a significance level $\alpha$, if $p$-value is less than $\alpha$, then we can conclude that there is statistically significant evidence strong enough to reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

This algorithm outlines a version of the bootstrap method. The choice of the test statistic is up to the user's discretion. Usually, the studentized $t$ statistic $T$ is used as the test statistic for comparing mean difference, which is given as follows.

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \tag{10}$$

If the $t$ ratio is chosen as the test statistic, then the bootstrap procedure will admit a slight modification, which then will output both the $p$-value and a $1 - \alpha$ level studentized confidence interval. Instead of counting, we

initialize an array to store the bootstrapped $t$ ratio in each iteration, which is calculated using the bootstrapped samples. After the loop, we sort the $t$ ratios in an ascending order, which then form the bootstrap cdf of the $t$ ratios. Then the confidence interval is given by

$$[\bar{X} - \bar{Y} - \hat{t}_{1-\alpha/2} \cdot \hat{SE}, \bar{X} - \bar{Y} - \hat{t}_{\alpha/2} \cdot \hat{SE}] \tag{11}$$

where $\hat{SE}$ is the pooled sample standard error given by

$$\hat{SE} = \sqrt{S_X^2/n + S_Y^2/m} \tag{12}$$

The corresponding $p$-value is defined as the largest $\alpha$ such that the confidence interval contains 0. Note that $p$-values computed this way is not exact, so the conclusion should be based on the confidence interval. This corresponds to a two-sided test. A one-sided test and a one-sided confidence interval can be derived similarly.

Additionally, if the test statistic is not the mean but some other statistic of which the form of its standard error may not be known, then a double loop will be required. The inner loop will estimate the standard deviation of the test statistic for each bootstrapped sample, based on which a $t$ ratio can be computed and thus a confidence interval can be derived in a similar fashion.

## D.5 Permutation Test

The permutation test implemented in `NLPStatTest` is the computational/randomization alternative to the exact sign test. It is used to test for consistent difference between two samples, which are assumed to be paired. In a sign test calibrated by the permutation, a relevant test statistic must be chosen in order to formulate the null hypothesis. The mean and the median are common choices.

The assumptions for running the permutation test are as follows:

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired.
2. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are independent.

The relevant hypothesis testing problem which can be calibrated by the permutation is for testing symmetric distribution around 0, or equivalently for the hypothesis that the median of the population is 0, given the underlying distribution is symmetric. The test procedure is given as follows.

**Algorithm 2** Paired Permutation Test

**Input:** $X, Y$, two paired samples
**Output:** $p$, permutation $p$-value

**procedure** SIGN TEST
  Calculate the difference $Z = X - Y$
  Calculate the observed test statistic of choice $\theta_{ob}(Z)$
  Given a large number $B$
  Given a counting number $C \leftarrow 0$
  **for** $b \in 1 : B$ **do**
    For each $Z_i$, change the sign with probability 0.5
    Calculate permutation test statistic $\theta_p(Z)$
    **if** $\theta_p \geq \theta_{ob}$ **then**
      $C \leftarrow C + 1$
  Calculate permutation $p$ value $= \frac{C+1}{B+1}$
  **return** $p$-value

As usual, given a significance level $\alpha$, if $p$-value is less than $\alpha$, then we can conclude that there is statistically significant evidence strong enough to reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

# E Appendix: Effect Size Estimators

This appendix provides definitions and formulae for effect size estimators implemented in `NLPStatTest`.

## E.1 Cohen's $d$ and Hedges' $g$

**Definition 11 (Standardized mean difference).** Consider two samples $\mathbf{X}_n$ and $\mathbf{Y}_m$, both **normally distributed** with mean $\mu_X$ and $\mu_Y$ and equal variance $\sigma^2$.

$$\delta = \frac{\mu_X - \mu_Y}{\sigma} \tag{13}$$

**Definition 12 (Cohen's $d$).** The Cohen's $d$ to measure the effect size of a test of two-sample mean difference is given by

$$d = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\hat{\sigma}} \tag{14}$$

where $\hat{\mu}$ and $\hat{\sigma}$ denote the sample mean and standard deviation.

The Cohen's $d$ is positively biased, implying an overestimation in small samples. To adjust this small bias, the Hedges' $g$ is thus proposed, multiplied with a correction factor $J$ which depends on the degree of freedom. The degree of freedom is the number of free parameters that can be estimated. In an unpaired two-sample testing scenario, the degree of freedom is the sizes of both samples minus 2. In a paired two-sample testing scenario, the degree of freedom is the number of pairs minus 1. The correction factor $J$ is given by

$$J(d.f.) = 1 - \frac{3}{4 \times d.f. - 1} \tag{15}$$

**Definition 13 (Hedges' $g$).**

$$g = J(d.f.) \cdot d \tag{16}$$

where $d$ is the Cohen's $d$ and $d.f.$ is the degree of freedom.

## E.2 Wilcoxon $r$

Wilcoxon $r$ is an effect size index for the Wilcoxon signed-rank test, calculated as $r = \frac{Z}{\sqrt{n}}$, where

$$Z = \frac{W - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum_{t \in T} t^3 - t}{48}}} \tag{17}$$

Here, $W$ is the test statistic for Wilcoxon signed rank test and $T$ is the set of tied ranks.

## E.3 Hodges-Lehmann Estimator

Hodges-Lehmann Estimator is an estimator for the median. Let $w_i = u_i - v_i$. The $HL$ estimator for one-sample testing is given by

$$HL = \text{median}\Big(\{(w_i + w_j)/2,\ i \neq j\}\Big) \tag{18}$$

# F Appendix: Power Analysis

Statistical power covaries with sample size, effect size and the significance level $\alpha$. In particular, power increases with larger sample size, effect size, and $\alpha$.

## F.1 Prospective Power Analysis

Prospective power analysis is used when planning a study (usually in clinical trials) in order to decide how many subjects are needed. In the NLP field, when one constructs or chooses a test corpus for evaluation, it will be beneficial to conduct this type of power analysis to determine how big a corpus needs to be in order to ensure that the significance test reaches the desired power level. `NLPStatTest` implements prospective power analysis based on Fundamentals of Biostatistics (Rosner, 2010).

## F.2 Retrospective Power Analysis

Retrospective power analysis is usually done after a significance test to determine the relation between sample size and power.

There are two scenarios associated with retrospective power analysis: When the values in $\{u_i - v_i\}$ are from a known distribution, one can use Monte Carlo simulation to directly simulate from this known distribution. To do this, one has to have an informed guess of the desired effect size (i.e., mean difference) via meta-analysis of previous studies. The algorithm is given in Algorithm 3.

**Algorithm 3** MC power simulation

**Input:** $T(\cdot)$, a statistical test; $\alpha$, sig. level
**Output:** $p$, power curve

**procedure** POWER SIMULATION
  Given a large number $B$
  Given a large number $N$ as sample size
  **for** $n \in 1 : N$ **do**
    Given a vector $\hat{p}$ to store $p$ values
    **for** $b \in 1 : B$ **do**
      Generate $n$ random samples under $H_1$: $\{X_i : i = 1, \cdots, n\}$
      Run test $T(X_1, \cdots, X_n, \alpha)$
      Obtain $p$ value and assign to $\hat{p}(b)$
    $p(n) \leftarrow \text{proportion}(\hat{p} < \alpha)$
  **return** $p$

When the distribution of the sample is unknown *a priori*, one can resample with replacement from the empirical distribution of the sample (a.k.a. the *bootstrap* method) to estimate the power. The algorithm is given in Algorithm 4.

**Algorithm 4** Bootstrap power simulation

**Input:** $X$,$Y$, data; $\delta$, effect size; $T(\cdot)$, a test; $\alpha$, sig. level
**Output:** $p$, power curve

**procedure** POWER SIMULATION
  Given a large number $B$
  Given a large number $N$ as sample size
  **for** $n \in 1 : N$ **do**
    Given a vector $\hat{p}$ to store $p$ values
    **for** $b \in 1 : B$ **do**
      Sample $X, Y$ of size $n$ with replacement, denoted by $X^b, Y^b$
      Shift $X^b$ or $Y^b$ by $\delta$ depending on the alternative
      Run test $T(X^b, Y^b, \alpha)$
      Obtain a $p$-value
    $p(n) \leftarrow \text{proportion}(\hat{p} < \alpha)$
  **return** $p$

# G Appendix: Sample Data Description

`NLPStatTest` package includes multiple sample data sets for users to experiment on.

| File name | Distribution | $H_0$ | $H_1$ | $\delta$ | $p$-value | Power |
|---|---|---|---|---|---|---|
| data-h0-equal-beta | beta | True | $=$ | 0 | high | low |

Table 4: Table of sample data and expected results