On the Statistical Significance Testing for Natural Language Processing

Haotian Zhu

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2020

Reading Committee:

Fei Xia

Gina-Anne Levow

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

On the Statistical Significance Testing for
Natural Language Processing

Haotian Zhu

Chair of the Supervisory Committee:
Professor Fei Xia
Linguistics

This thesis explores and compares statistical significance tests frequently used in comparing Natural Language Processing (NLP) system performance in several aspects. We begin by establishing the fundamentals of the NLP system performance comparison and formulating it into four major tasks specific to NLP. Each statistical significance test is explained in great detail with its assumptions explicated and testing procedure outlined. We stress the importance of verifying test assumptions before conducting a test. In addition, we examine the effect size and statistical power and discuss their significance in the statistical significance testing in NLP. By considering potential dependencies within a test set, the block bootstrap is introduced and employed to calibrate the statistical significance testing for comparing performance of two systems on average. Four case studies with both simulated and real data, of which the complexity of data dependency varies, are presented to illustrate the process of properly using a statistical significance test in comparing NLP system performance under different settings. We then proceed to discussion from different perspectives, with some open issues such as cross-domain comparison and the violation of *i.i.d.* assumption, which expects further studies. In conclusion, this thesis advocates the proper use of statistical significance testing in comparing NLP system performance and the reporting of the comparison results in more transparency and completeness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF NOTATION

$\mathbf{X}$: a random variable (a sample)

μ_X: population mean of $\mathbf{X}$

σ_X: population standard deviation of $\mathbf{X}$

$\bar{\mathbf{X}}$: the sample mean of $\mathbf{X}$

$\hat{\sigma}_X$: the sample standard deviation of $\mathbf{X}$

X_I: an observation in a sample

$D.F.$: degree of freedom

$\perp\!\!\!\perp$: statistically independent

H_0: the null hypothesis

H_1: the alternative hypothesis

$P_0(\cdot)$: the conditional probability under the null hypothesis

$P_1(\cdot)$: the conditional probability under the alternative hypothesis

α: the significance level or Type I error

β: the Type II error

F_X: the cumulative distribution function of $\mathbf{X}$

$\mathbf{X} \sim \mathbf{Y}$: $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution

$A1$: Assumption 1

$T1$: Task 1

$\text{RANK}(X_I)$: the rank of observation X_i

$\mathbb{I}$: the indicator function

ACKNOWLEDGMENTS

I would like to first extend my gratitude to my advisor Professor Fei Xia for her invaluable and meticulous guidance throughout the way. It is under her auspices that, from a quixotic and immature idea, through numerous hour-long discussions via Skype, this thesis is made possible and able to come to its fullness. I am also grateful to Professor Gina-Anne Levow, for her indispensable suggestions punctually provided in carefully annotated documents. I would like to express my appreciation to Professor Emily Bender for her understanding and support when the path before me was obfuscated. I would like to appreciate the department of linguistics for efficient and magical, I would say, administrative assistance. During these three years of my graduate life, solitude and insomnia have always accompanied me, insidiously festering in invisibility, but even in the most desperate time of all there is always a scintilla of glistening flare. That flare comes from unwavering friendship and kinship. To my friends Fan Wang, I will always remember the time and felicity we have shared together. Lastly and most important of all, I thank my parents, for their undivided and unconditional love and support, for their unlimited patience and tolerance for my fractious temperaments, and hence, to them, I am forever indebted.

DEDICATION

In dedication to my parents, friends and those who've offered me time and company.

Chapter 1
INTRODUCTION

The system performance comparison among NLP systems relies largely on the statistical significance test as a standard empirical approach to ensuring that experimental results generated by the proposed system are not due to happenstance but a real effect brought by the proposed system. Despite its importance in empirical studies, the statistical significance testing has received relatively limited attention in the field of NLP. Smucker, Allan, and Carterette (2007), Bisani and Ney (2004) and Koehn (2004) focused on significance testing applied to specific NLP tasks such as information retrieval, automated speech recognition and machine translation; Dror et al. (2018) and Berg-Kirkpatrick, Burkett, and Klein (2012) drew comparisons of statistical tests in a wider range of NLP tasks.

Given a proper choice of evaluation metric, the system performance comparison task usually considers the realized metric difference between two systems as the discriminatory factor in order to determine superiority of system performance on a test set. To simplify, in this thesis, we consider a baseline system and a proposed system for competition. The major goal is to draw a statistically sound conclusion as to whether or not the newly proposed system performs better than the baseline system on the given test set:

$$H_0 : \textit{ the proposed is better than the baseline.} \tag{1.1}$$

Put in concrete statistical terms, the comparison could be done with respect to average performance, a common hypothesis in statistical hypothesis testing which is phrased with respect to the population mean of the sample–whether or not the two samples are different on average, in a two-sample testing. Alternatively, due to the sensitivity of the sample mean to the presence of outliers, a more robust performance comparison based on the median can be used as well. Some may also suggest a comparison based on the probability distributions

of the evaluation metric, where the relevant conclusion is drawn with respect to whether the two systems produce identically distributed values of the chosen evaluation metric.

In each comparison task, there exist a number of applicable statistical tests, which must be conducted in accordance with assumptions on the data. In testing the population mean difference, the student t test is a frequently used and classic test which relies on many statistical assumptions. Usually such statistical assumptions tend to be unrealistic for real-world NLP data. For example, many tests rely upon the assumption that the sample contains independent and identically distributed ($i.i.d.$) observations, which may not be satisfied and in essence nullify the applicability of the tests. However, before conducting statistical significance tests, relevant studies have dedicate limited attention to verifying the assumptions, even though it is of vital importance.

Many previous studies on the subject of statistical testing in NLP tend to refrain from formulating the comparison task in statistical terminologies or even under a statistical framework. To better understand and present the statistical assumptions of each test and the comparison task itself, this thesis will start with definitions (Chapter 2) for the system performance comparison task in NLP and frequently used statistical terms for clarification and specification. Then, this thesis will present and explain some statistical assumptions shared by many statistical significance tests. Since a relatively simplified comparison task which only considers one test set is discussed in this thesis, assumptions made throughout this thesis will be explicated.

It might be standard practice for some to base their decisions solely on the p-value obtained from a significance test. The p-value, which measures the probability of having a test statistic that is as 'extreme' as the observed one, has been at the center of debate in some fields of empirical studies for many years, due to its misuse and misleading interpretation. Many empirical study papers in NLP which proposed novel systems usually report the obtained p-values and then conclude on statistically significant superiority of the proposed system to the baseline. Nevertheless, such significance due to small p-values is expected if the sample size of the test set is large, meaning that a test based on a large sample will also

produce significant results. In such case, a post hoc power analysis and effect size analysis will be required to determine how statistically powerful the chosen significance test is and what the effect size of this proposed system. Statistical power measures the probability that if your hypothesis is false, how likely the test is going to reject the hypothesis; the effect size, independent of the size of the test set and the test, measures the actual magnitude of the performance difference between the proposed and the baseline systems. In this thesis, we will offer two ways of estimating statistical power of a significance test and define some indices of effect size which are relevant in NLP system performance comparison.

As aforementioned, each significance test has its own assumptions and a set of hypotheses to test. The major differentiating factor in assumptions in statistical significance tests is in terms of parametricity: whether or not a statistical significance test puts an assumption on the probability distribution of the sample. If so, then this test is said to be parametric; if not, the test is nonparametric. Parametric and nonparametric tests are two main families of statistical tests, and tests in each family have different hypotheses. The student t test is a famous parametric test which assumes normally distributed sample(s) and aims to test mean difference (two sample testing) or mean equality; the Wilcoxon signed-rank test is a rank-based nonparametric test which aims to test for symmetry of the probability distribution around 0.

In the family of nonparametric tests, statistical significance tests calibrated by randomization methods also receive popularity among researchers due to the rising computational power and their more relaxed statistical assumptions, such as the bootstrap method used in ASR system evaluation (Bisani and Ney, 2004). The permutation and the bootstrap are two frequently used methods to calibrate statistical significance tests. In this thesis, we will present a selection of parametric and nonparametric tests, with their assumptions, hypotheses, testing procedure and their usage in NLP system performance comparison elaborated, Also, the permutation, the bootstrap and their extensions will be discussed in NLP system performance comparison tasks.

It has been mentioned in previous work that the issues related to dependencies within

test set are prevalent and violate the assumptions held by some statistical tests (Yeh, 2000). To accommodate the dependencies existing in NLP data such as dependent sentences in the same article in a machine translation task, we propose a test based on the block bootstrap (or the cluster bootstrap) which takes such dependency into account.

After proposing a hypothesis and choosing an appropriate test, the testing procedure is usually done under the auspices of computer softwares such as *Python* and *R*. These softwares will output the test result including the p-value. It has been advised by many associations (including the American Statistical Association) and journals that reporting only the p-values will not be sufficient in published researches. Thus, to facilitate the reporting of a more convincing and statistically sound test result, the estimation of statistical power of the chosen test and the computation of the effect size of the experiment will hence be needed in post hoc analysis. We will hence, using both simulated and real data, demonstrate the proper use of statistical significance tests in multiple illustrative examples where different data types common seen in NLP tasks appear and in each case study compare the statistical power of many tests in post hoc exploratory analyses.

This thesis is organized as follows. To start, Chapter 2 presents preliminary definitions used in this thesis and assumptions shared by statistical tests. The foundation and a general procedure of statistical hypothesis testing are also briefly discussed in this chapter. In drawing conclusions, the use of p-values, effect size and power of tests will be addressed as well. Various types of hypotheses are also given in Chapter 2 to facilitate the discussion on the application to NLP tasks in later chapters.

Chapter 3 reviews previous work on significance tests in NLP from methodological and empirical angles. In methodological papers, we highlight the tests the authors introduce and considerations they make. In empirical papers, we discuss the authors' assumptions on the test set and their use or misuse of the tests.

Chapter 4 is the major chapter which introduces all relevant tests for NLP system comparison, starting with the way to verify test assumptions, parametric tests, nonparametric tests and tests based on compute-intensive methods such as the permutation and the bootstrap.

Considerations on employing a statistical test and on reporting test results are explored in Chapter 5. Illustrative case and simulation studies are also given, where we demonstrate in concrete examples the proper way of conducting a test to compare system performance. Examples include numerical and categorical data, both in independent and dependent settings.

Chapter 6 reexamines findings in this thesis and previous work in regards to their limitations, and Chapter 7 serves as the conclusion chapter and prospect for future work.

Chapter 2

PRELIMINARIES AND ASSUMPTIONS

Before going into details about the statistical significance testing itself, it is necessary to explain some recurring statistical terminologies and definitions used in this thesis first. In this chapter, we will present definitions and assumptions used in this paper and preliminaries of the statistical significance test, where a general hypothesis testing procedure is given. In addition, this section will also discuss the proper use and interpretation of the p-value, a commonly adopted measure of significance and factor of making decisions. The use of effect size and power analysis in empirical studies will be introduced in this chapter. We will also discuss frequently seen data types in NLP system comparison tasks.

There are four major comparison tasks in NLP system performance comparison. The first and most frequently used one is to compare the population means of two samples (Subsection 2.8.1). The second is to compare the probability distributions of two samples (Subsection 2.8.2). The third task is to determine whether two samples are independent or not (Subsection 2.8.3), and the last task is to test for equality of medians of two samples or equivalently testing for symmetry (Subsection 2.8.5). A more general task can also be considered, which is to test for a certain statistic or parameter of a statistical model (Subsection 2.8.4). We will provide detailed definitions for the four tasks in this chapter. Finally, a simplified table of tests will be given to navigate future chapters.

2.1 Definitions

In this section, we define the terms used in this thesis. For a system performance evaluation task, we consider a simplified case of comparing two systems, a proposed system and a baseline system. The goal is to determine statistically whether or not the proposed system

outperforms the baseline system with respect to the same test set and a relevant evaluation metric. We first define what we mean by a test instance.

Definition 1 (**Test instance).** A test instance, denoted by t, is a system input of size 1 in a data type (numerical, categorical, etc.) required by the NLP system of interest. A test instance must be independent from the training data used to train the system and must be held out until the testing procedure is completed.

Definition 2 (**Evaluation unit).** An evaluation unit $e = \{t_i, i = 1, \cdots, n\}$ is a set of test instances upon which an evaluation metric can be defined.

Definition 3 (**Test set).** A test set, denoted by $T = \{e_j, j = 1, \cdots, m\}$, is a collection of evaluation units.

Definition 4 (**Evaluation metric).** Given a system A, the evaluation metric E for the NLP task is a function of the output of A given the test set $T = \{e_j, j = 1 \cdots, m\}$ which transforms the output to a numerical value, denoted by $E_A(e_j)$, for some evaluation unit e_j.

An evaluation unit is an intermediate level between the system output and the values of evaluation metric. Conventionally, given an NLP system, we have a test set, which is then thrown into the system and a system output is produced. Given the system output, we then evaluate the system output using an appropriate evaluation metric. The evaluation metric can be defined on a single instance of the system output, for example a sentence-level $BLEU$ score calculated based on a single translated sentence; alternatively, the evaluation metric can also be defined on a set of system output instances, for example an accuracy calculated based on a set of sentences composed of multiple tokens. The notion of an evaluation unit is not clearly defined in previous studies of NLP system performance comparison, but it is beneficial to make such distinction and give a precise definition of an evaluation unit. An evaluation unit can be a set containing only one test instance, in the case where the evaluation metric is defined with respect to a single test instance. For example, in machine translation, the evaluation metric $BLEU$ is defined with respect to a single sentence, in which case a

test instance of a single sentence is an evaluation unit, and a machine translation system requires a single sentence as an input. An evaluation unit can also contain multiple test instances. For example, the performance of a POS tagger is usually evaluated by accuracy, which is calculated by the number of correctly tagged words over the total number of words. However, the POS tagger may only require a single word as a system input. In this case, a test instance is a single word; an evaluation unit is a set of words, from which one can calculate accuracy; a test set is a set of multiple sets of words.

Note that usually the system output is a one-to-one function of the test set, meaning each test instance corresponds to one and only one output instance. Therefore, for simplicity, we ignore the notation for the system output and regard the evaluation metric directly as the function of the evaluation unit in the test set.

The following definitions are statistical definitions used in this thesis. Many methodological papers which broach the assumptions of statistical tests mentioned the notion of independence. In a statistical sense, independent samples are those samples of which the joint probabilities are product of their corresponding marginal probabilities. Note that this notion of statistical independence is not the same as random sampling, which connotes a sampling procedure producing statistically independent and identically distributed samples.

Definition 5 (**Statistical independence).** Consider two discrete random variables X and Y. X and Y are independent, denoted by $X \perp\!\!\!\perp Y$, if and only if $\forall x, y$

$$P(X = x, Y = y) = P(X = x)P(Y = y) \tag{2.1}$$

If X and Y are real-valued continuous random variables, then $X \perp\!\!\!\perp Y$ if and only if $\forall x, y$

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \tag{2.2}$$

or equivalently, if the joint probability density function exists, $X \perp\!\!\!\perp Y$ if and only if $\forall x, y$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \tag{2.3}$$

Definition 6 (**Random sample).** A sample $X_1, \cdots, X_n$ is random if all X_i's are sampled from the population, which can be finite or infinite, randomly and independently, where the probability of choosing one does not affect the probability of choosing another.

The notion of random sampling is addressed in many statistics books. Efron and Tibshirani (1993) defined a random sample to be a sample of size n which is a collection of units $u_1, \cdots, u_n$ drawn from a population $\mathbb{U}$ (suppose the population is finite of size $n \leq m$) at random. The process of random sampling is ideally done by a random number generator which generates a random integer number i between 1 and m with probability $1/m$, and then the unit u_i is selected. A random sample drawn with replacement from the population satisfies the independent and identically distributed condition; a random sample drawn without replacement from the population, however, does not satisfy such condition but instead exchangeability.

Definition 7 (**Identical distribution).** Two random variables X and Y are identically distributed, denoted by $X \sim Y$, if $\forall x \in D$, $P\big(X \leq x\big) = P\big(Y \leq x\big)$, where D is the domain of X and Y. Equivalently, X and Y are identically distributed if they have the same cumulative distribution function (CDF).

Definition 8 (**Independence).** A sample $X_1, \cdots, X_n$ is called *i.i.d.*, i.e., independent and identically distributed if $X_1, \cdots, X_n$ are mutually independent and all X_i's follow the same distribution.

Definition 9 (**i.i.d.).** A sample $X_1, \cdots, X_n$ is called *i.i.d.*, i.e., independent and identically distributed if $X_1, \cdots, X_n$ are mutually independent and all X_i's follow the same distribution.

Definition 10 (**Exchangeability).** An exchangeable sequence of random variables is a sequence of random variables $X_1, X_2, X_3, \cdots, X_n$ such that for any permutation π of indices $1, 2, 3, \cdots$, the joint probability distribution of the permuted sequence

$$X_{\pi(1)}, X_{\pi(2)}, X_{\pi(3)}, \cdots$$

is the same as the joint probability distribution of the original sequence. A sequence of $i.i.d.$ random variables is exchangeable but not necessarily vice versa.

Definition 11 (**Significance level).** The significance level of a significance test is a pre-determined quantity of subjective choice $\alpha \in (0, 1)$ such that it measures the degree of risk one would like to suffer from. α is also called the **Type I Error**, which is the probability that when the null H_0 is true, the null hypothesis is rejected by the test.

$$\alpha = P_0\big(H_0 \; is \; rejected\big) \tag{2.4}$$

Definition 12 (**Type II error).** The Type II error of a significance test is the probability that under the alternative H_1, the alternative is rejected by the test, which is usually denoted by β.

$$\beta = P_1\big(H_1 \; is \; rejected\big) \tag{2.5}$$

Definition 13 (**Power).** The power of a statistical significance test measures the probability that under the alternative, the null is correctly rejected, or the alternative is accepted by the test. The power of a test is simply $1 - \beta$.

Definition 14 (P**-value).** The p-value of a significance test is the probability that under the null, the test statistic is as 'extreme' as the observed one, where the term 'extreme' is conditional on the rejection criterion–that is we reject for large values of the test statistic or for small values. Suppose the test statistic is $\theta(X)$ and we reject for large values of θ. The p-value is given by

$$p = P_0\big(\theta(X) \geq \theta_{observed}\big) \tag{2.6}$$

Definition 15 (**Stochastic dominance).** Consider two continuous random variables A and B, defined on the same support $D = (-\infty, \infty)$. A (first order) stochastically dominates B if

$$\forall x \in D, \; P\big(A \geq x\big) \geq P\big(B \geq x\big) \tag{2.7}$$

and for some $x \in D$,

$$P\big(A \geq x\big) > P\big(B \geq x\big) \tag{2.8}$$

Let $F_A(x)$ and $F_B(x)$ be the cumulative distribution functions of A and B. Equivalently, A (first order) stochastically dominates B if

$$\forall x \in D,\ F_A(x) \leq F_B(x) \tag{2.9}$$

and the strict inequality holds for some $x \in D$.

2.2 Assumptions

In this section, we provide general statements which are presupposed in this thesis. These assumptions will not be tested but rather assumed throughout this thesis.

Assumption 1. *We assume that the evaluation metric is numerical/cardinal, and thus basic arithmetic operations such as addition and subtraction can be applied.*

Assumption 2. *We assume that the relevant evaluation metric truthfully represents the 'goodness' of the system performance. Then the decision rule is to prefer the system with higher values of evaluation metric E.*

Then it can be seen that the essential goal is to compare the magnitude of evaluation metric, through which we can thus draw conclusions on the performance of the systems. To formulate this system performance comparison task into a proper hypothesis testing problem, the following important assumption has to be made with respect to the value of the evaluation metric $E_A(e_j)$.

Assumption 3. *We assume that given a system and a relevant evaluation metric, the generation of values of the evaluation metric is* ***probabilistic****.*

That is to say that we place system performance comparison in the probabilistic theoretical construct and we view the value of the evaluation metric of a particular system as a random variable which follows some probabilistic distribution.

There general assumptions may seem trivial and taken for granted, and they are often omitted by many papers which actually rely on them. Thus, it is necessary to accentuate their existence and importance.

Assumption 4. *We assume that the dependency, if any, which exists among the evaluation units, is preserved after applying the evaluation metric.*

This means that the way in which the evaluation units are correlated with each other is consistent with the way in which their corresponding values of evaluation metric are correlated. Thus, we only need to direct our attention to the correlation among values of the evaluation metric in consideration.

Statistical tests have their corresponding assumptions in order to be performed. It is important to verify the assumptions of a test before conducting the test. The following assumptions are shared among many statistical tests. Consider two random samples $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{Y} = (Y_1, \cdots, Y_m)$, where n and m are natural numbers denoting sample sizes of $\mathbf{X}$ and $\mathbf{Y}$.

Assumption 5 (**Independent samples**). *The sample* $\mathbf{X}$ *is independent of the sample* $\mathbf{Y}$.

Assumption 6 (**Exchangeable samples**). *The sample* $\mathbf{X}$ *and sample* $\mathbf{Y}$ *are both exchangeable, as described in Definition 10.*

Assumption 7 (**Independent observations**). *The sample* $\mathbf{X} = (X_1, \cdots, X_n)$ *contains independent observations* $X_1, \cdots, X_n$.

Assumption 8 (**Identically distributed samples**). *The sample* $\mathbf{X} = (X_1, \cdots, X_n)$ *contains identically distributed observations* $X_1, \cdots, X_n$.

Assumption 9 (**Normality**). *The sample* $\mathbf{X}$ *is normally distributed.*

Assumption 10 (**Homoscedasticity**). *The variances of* X_i *and* Y_j *are equal, i.e.,* $\mathrm{Var}(X_i) = \mathrm{Var}(Y_j)$.

Assumption 11 (**Representative samples**). *The sample* $\mathbf{X}$ *and sample* $\mathbf{Y}$ *are representative of their corresponding populations.*

Assumption 12 (**Linear association**). *The sample* $\mathbf{X}$ *and sample* $\mathbf{Y}$ *are linearly associated, meaning that given two coefficients* α *and* β*, the association can be (approximately) described by* $\mathbf{Y} = \alpha\mathbf{X} + \beta$.

Assumption 13 (**Monotone association**). *The sample* $\mathbf{X}$ *and sample* $\mathbf{Y}$ *form a monotone association if the directions of changes in both variables are the same.*

Note that a sample is a collection of data points, and each data point is a random variable. Thus a sample is a vector of random variables. We divide the assumption of *i.i.d.* into two separate part because some tests only assume one of the two.

2.3 Data type

This section will introduce different types of variables. The most common type of variable, or data, is called the **numerical** variable, which, videlicet, consists of quantitative variables of which the numerical values are of importance. For example, a $BLEU$ score is a numerical variable. Under numerical variables, there two subtypes, one of which is **continuous** variables and another of which is **discrete** variables. These two subtypes denote the specific probabilistic distribution according to which the data are distributed. Some statistical tests require the data to be continuous.

Another commonly seen type of variable is called the **categorical** variable, which take categories or labels as its value. For example, the system output for a POS tagger is categorical, and we evaluate the system output using confusion matrix or accuracy.

There is also a type of variables called the **ordinal** variable, where there is a clear order. For example, the income levels of low, middle and high are ordinal. Ordinal data can be either categorical or numerical.

In a two-sample problem, it is possible that the two samples are independent of each other or the two samples are dependent, the latter of which is usually referred to as repeated measures or matched pairs, which will be frequently encountered later on. In performance comparison tasks of NLP systems, it is usually the case that we are facing 2-sample data in form of matched pairs: two systems evaluated on the same test set. Sometimes, researchers might be also interested in comparing multiple systems.

2.4 Statistical hypothesis testing procedure

Classical statistical hypothesis testing is a scientific framework upon which researchers rely to draw statistically sensical conclusions. This framework consists of **a null hypothesis** H_0 and **an alternative hypothesis** H_1. Usually the research hypothesis in consideration is regarded as H_1, and H_0 is phrased as the opposite of H_1. Hypothesis testing often uses the testing procedure outlined in Table 2.1.

Table 2.1: General testing procedure

- Propose a research question.
- Phrase the research question (quantitatively or qualitatively) into the alternative hypothesis and formulate a null hypothesis contrary to the alternative.
- Design a test (analytically or computationally) which assumes that the null hypothesis is true.
- Collect data in accordance with the test and do not examine the data before conducting the test (after choosing a test, one can use the data to verify assumptions).
- Conduct the test using the collected data.
- Based on the test result, reject the null or fail to reject it and draw a conclusion accordingly.

The statistical hypothesis testing is based on the idea of proof by contradiction. One begins by assuming H_0 is true and attempts to derive a contradiction, for which one can reject the null hypothesis. However, the hypothesis test is fundamentally premised upon

probabilistic model, where it cannot categorically prove or disprove the truth condition of the hypothesis in consideration; rather it is designed to show that there exists statistically significant evidence which is strong enough to reject the null hypothesis H_0. It is often a misconception or misuse that the hypothesis test can result in a definite conclusion which accepts the null or alternative hypothesis. The only statistically permissible conclusion to make is either to reject the null hypothesis or to fail to reject the null hypothesis.

The statistical hypothesis test is more like an experimental design in fact. Note that in the statistical hypothesis test, it is required that one must not design a test according to what the data look like. Some researchers do so by conducting a variety of statistical tests on the same set of data and choose the one with the highest significance to report. This is called 'p-hacking', and it is best that it is avoided.

2.5 A note on the p-value

The American Statistical Association has issued a statement (Wasserstein and Lazar, 2016) which includes six principles (Table 2.2) underpinning the appropriate use, interpretation of the p-value and decisions made based on the p-value.

The first two bullets in this statement dictates the function of the p-values, which is to measure the compatibility of the data with a proposed statistical model. It does not determine the degree to which the hypothesis is true or false. The third bullet advises that conclusions or decisions not be based on the p-values alone, which is however usually done by many researches by just looking at the significance level and the obtained p-values. Such sole dependence on the p-values is likely to bring about ramifications as in NLP, the sample size of the test set is usually large; with large test sets, the p-values will always be small or even small enough to indicate statistical significance (Lin, Lucas, and Shmueli, 2013). The fourth bullet is suggesting that to make proper statistical inference, reporting in completeness and transparency of the test results is required. This recommended practice is also ignored in most experimental papers in NLP, where the p-values are the only numbers reported in the system evaluation results and no information on the effect size or power is provided to allow

Table 2.2: The statement on p-values

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

for reproducing the work. The fifth bullet accentuates the importance of the effect size and the limitations of the p-values, which gives rise to the necessity of the consideration of the effect size of a test. These principles, however, do not necessarily discourage the use of p-values in empirical research or devalue them, but advocate the reporting of test results in a more transparent and sufficient way, not just with p-values themselves.

2.6 Effect size

Cohen (1994) in his book *Statistical Power Analysis for the Behavioral Sciences* examined the use of effect size in empirical research and defined it as the degree to which the 'phenomenon' is present in the population, or the degree to which the null hypothesis is false. It is a common

misunderstanding that if the obtained p-value is less than 0.05, then the effect must be big enough. Cohen suggested that p-values are probability values, which derive from the effect size and the sample size. For any non-zero effect size, if the sample size is large enough, then the resulting p-value could suggest significance, which is misleading.

Moreover, Cohen claimed that when the null hypothesis is false, the effect size is some specific nonzero value on the population level. The larger the effect size is, the greater the degree to which the 'phenomenon' is present in the population. The term 'phenomenon' in his notion refers to, for example, the difference in the parameters in consideration as a result of the experiment group, compared to the control group, in an experimental study (Cohen, 1988). Also, note that the effect size usually appears in calculating the required sample size of the test and in the report of test results; the effect size is not for comparing tests or a metric for choosing which test to use. It is suggested that the effect size is reported in a study along with the p-value.

As an extended study on meta-analysis, Cooper, Hedges, and Valentine (2009) claimed that effect size is appropriate when the index quantifies the relationship between two variables or the difference between two groups. They also mentioned that there exist four properties the effect size should have. First of all, the effect size should not depend on factors associated with the design of a study, such as the sample size or involved covariates. Additionally, the effect size should convey certain meaningfulness and interpretability. The effect size, as suggested by the authors, should also be computed from reports of a published research, and lastly, the effect size's sampling distribution should be known or estimated in order to compute variances and confidence intervals.

In NLP system performance comparison tasks, many researchers are interested in comparing population means of values of the chosen evaluation metric between a proposed system and a baseline system, aiming to make inference on the average performative superiority. The commonly used index of effect size for mean comparison is the standardized mean difference, as proposed by Cohen. Cooper, Hedges, and Valentine (2009) however claimed that the Cohen's d is biased and may overestimate the absolute value of the true effect size. Thus,

to adjust this bias, the Hedges' g is proposed. We will introduce both indices here, in the case where the two samples are paired and thus dependent.

2.6.1 Cohen's d

The effect size is considered an unknown number on the population level, denoted by δ by Cooper, Hedges, and Valentine (2009). The population level effect size of the standardized mean difference is defined by

Definition 16 (**Standardized mean difference).** Consider two samples $\mathbf{X}_n$ and $\mathbf{Y}_m$, both normally distributed with mean μ_X and μ_Y and equal variance σ^2.

$$\delta = \frac{\mu_X - \mu_Y}{\sigma} \tag{2.10}$$

The standardized mean difference is a population level parameter, which is assumed to be unknown and fixed. The Cohen's d is an appropriate estimate for the standardized mean difference.

Definition 17 (**Cohen's** d**).** The Cohen's d to measure the effect size of a test of two-sample mean difference in a parametric test is given by

$$d = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\hat{\sigma}} \tag{2.11}$$

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

However, when the assumed two normal distributions are not equally variable, then the Cohen's d is calculated by the following using the pooled standard deviation.

$$d = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{(\hat{\sigma}_X^2 + \hat{\sigma}_Y^2)/2}} \tag{2.12}$$

Depending on which sample is placed first in the numerator, the effect size can be negative or positive. To avoid this directionality, one can compute the absolute value of the difference and then divide it by the pooled standard deviation.

Cohen also proposed a qualitative interpretation of the Cohen's d as in Table 2.3. It is possible that the test produces a statistically significant result (p-value less than α) but the effect size is small, indicating a trivial difference. Therefore, it might be misleading to direct all attention to the p-value and ignore the effect size. In Table 2.3, if a Cohen's d is below 0.2, then the effect size is trivial and negligible. If between 0.2 and 0.5, then a small to medium effect is indicated. If between 0.5 and 0.8, a medium to large effect size is implied. If a Cohen's d is over 0.8, then the effect size is estimated to be large. It is obvious that d ranges from 0 to infinity.

Table 2.3: Cohen's d effect size

d	effect
0.2	small
0.5	medium
0.8	large

Instead of point estimation, we can also compute the variance of the Cohen's d in order to obtain a confidence interval. Denote the correlation between pairs of observations in the samples by r. Given n pairs of observations in a paired two-sample testing scenario, the variance of the Cohen's d is given by

$$V_d = \left(\frac{1}{n} + \frac{d^2}{2n}\right) 2(1 - r) \tag{2.13}$$

Then a 95% confidence interval based on a normal distribution is given by

$$[d - 1.96 \times V_d, d + 1.96 \times V_d] \tag{2.14}$$

2.6.2 Hedges' g

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

$$J(d.f.) = 1 - \frac{3}{4 \times d.f. - 1} \tag{2.15}$$

Definition 18 (**Hedges'** g)**.**

$$g = J(d.f.) \cdot d \tag{2.16}$$

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

The choice of indices between d and h is dependent on the presence of knowledge of the sample size in each sample. Cooper et. al. suggested that unless the sample size is less than 10, the two indices of effect size display no practical difference (Cooper, Hedges, and Valentine, 2009).

2.6.3 Common language effect size

To facilitate understanding of the effect size for audience outside the statistical academia, McGraw and Wong (1992) proposed an effect size statistic named the common language statistic (CL), which is defined as the probability of obtaining a difference score greater than zero in the distribution.

Definition 19 (**Common language statistic** CL)**.** Consider the sample difference $Z = X - Y$, where we assume that the mean of X is greater than that of Y.

$$CL = P\big(Z > 0\big) \tag{2.17}$$

The common language effect size assumes that both X and Y are normally distributed and have equal variances. Thus, the difference Z is also normal with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. The probability of Z being greater than 0 can be calculated by a z-score.

McGraw and Wong also gave the adjusted variance of Z for correlated samples, which is

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y \tag{2.18}$$

where r is the Pearson correlation.

In absence of normality and homoscedasticity, the authors conducted experiments and concluded that the violation of normality assumption alone does not seriously undermine the estimation of CL using the standard normal distribution; joint violation of both assumptions however does create a larger estimation error, but the authors believed that since the maximum error is always small, in absence of the assumptions, CL can still bear meaningful results.

2.6.4 Other indices and computation in R

The introduced Cohen's d and Hedges' g are used for parametric tests such as the student t test which assumes normality. For other nonparametric tests such as the Wilcoxon signed-rank test and the Mann-Whitney U test, other indices of effect size are appropriate.

$$r = \frac{Z}{\sqrt{n_{pair}}} \tag{2.19}$$

The index r is appropriate for the Wilcoxon signed-rank test and the Mann-Whitney U test, where Z is the normalized U statistic computed in the Mann-Whitney test (the test in R will have this output). In addition, the Cramer's V index is used for the Pearson chi-square test for categorical data.

The *rcompanion* package in R covers most of the indices of effect size.

2.7 Power analysis

In this section, we will investigate methods to simulate the statistical power of a test. Under the statistical hypothesis testing framework, there are two types of errors: the Type I error and the Type II error. The Type I error (α) is the probability under the null that the null is rejected by the test, which is also called the false positive. Type I error can be easily

controlled at a desired level by setting the significance level α, conventionally at 0.05 or 0.01, depending on the degree of risk the researchers wish to suffer. The Type II error (β) is the probability that the null is rejected when the null is false, which is called the false negative, conventionally set at 0.8. As defined earlier, the statistical power is the probability under the alternative that the null hypothesis is correctly rejected, which equals $1 - \beta$. Power is associated with sample size, effect size and α. Generally, as sample size or effect size grows to infinity, power of a test tends to 1. Given a fixed significance level, larger statistical power indicates more accuracy of the result and less error in the inference (Perugini, Gallucci, and Costantini, 2018). Power analysis is usually used prior to the actual hypothesis testing in order to determine the minimally needed sample size to achieve either a desired level of power or effect size.

There are different types of power analysis (Faul et al., 2007). In planning the system performance comparison task, the power analysis can provide useful information on the needed sample size of the test set or the desired effect size. Some also use the power analysis in retrospect in order to understand an existing study by computing the observed power. Prospective power analysis (in study planning and before data collection) and retrospective power analysis are not identical, and some researchers advise against the use of retrospective power analysis (Perugini, Gallucci, and Costantini, 2018; Zumbo and Hubley, 1998; Faul et al., 2007). One of the major criticisms given is that retrospective power analysis usually regards the estimated effect size and the sample size as the population effect size and population sample size, whereas the estimates are biased. Unfortunately, neither the population effect size nor the sample size is unlikely to be obtained. A post hoc power analysis, which is often conflated with the retrospective power analysis, intends to compute the power of a published study with the given significance level, sample size and a specified population effect size.

The computation of statistical power of a test can be done manually in many tests (the student t test). One can also resort to simulation and the bootstrap to estimate statistical power. Here we introduce two computational approaches to estimating statistical power of a

study. If the distribution of the sample, both under the null and the alternative, is known, we can use Monte Carlo simulation to directly simulate data points from the known distribution and conduct the test accordingly. Without any knowledge of the distribution but with only the actual sample in hand, the bootstrap can be used to repeatedly draw samples from the original sample where the original sample is regarded as a surrogate of the true population. Note that the two methods proposed in the thesis aim to estimate statistical power in a post hoc fashion, which implies that the population effect size is specified.

Monte Carlo simulation

With known distributions of the samples, the definition of statistical power suggests a computational way of simulation as specified in Algorithm 5 in Appendix B. The power simulation procedure utilizes Monte Carlo simulation, where we vary the sample size in each iteration. In the second loop, we generate random samples under H_1, run the statistical test of interest and obtain the p value. Then we compute the fraction of the p values which are less than α and obtain the simulated power at a particular sample size n, or alternatively, we can count how many times the test rejects the null hypothesis. This algorithm will give a power curve with different sample sizes.

Bootstrap power estimation

Given the samples without knowing their exact distributions, we could use the bootstrap method (Collings and Hamilton, 1988) to resample the data to create bootstrapped samples under the alternative hypothesis, conduct the relevant statistical test and then calculate the proportion of rejections of the null hypothesis. The outline of bootstrap power simulation is given in Algorithm 6 in Appendix B.

Collings and Hamilton (1988) also suggested an alternative method of estimating power using the bootstrap for a two-sample problem, which is called the average $X\&Y$. This method first computes the power for both samples X and Y by invoking Algorithm 6 with a slight modification that the two samples compared in this algorithm come from the same

empirical distribution (either from X or Y). After estimating the power for both X and Y, we compute the following quantity.

$$\frac{n\hat{\beta}_X + m\hat{\beta}_Y}{n+m} \tag{2.20}$$

where n, m are the sample sizes of X and Y. This method estimates the statistical power of a test with fixed significance level, fixed sample size and fixed effect size (which comes into play in shifting one of the samples by δ).

In NLP system performance comparison, the bootstrap power estimation relatively displays a greater use since it is rare that the exact distributions of the given sample are known. In prospective power analysis, what we usually have in hand is the baseline system to compare our proposed system with and a held-out test set, meaning that the sample size is fixed. By generating output from the baseline system using the test set and evaluating the baseline output with an appropriate evaluation metric, we now have a pilot sample. It is possible to do the following with the pilot sample. Firstly, given a desired significance level and an alternative hypothesis, we can use the bootstrap power estimation method to compute the effect size that can be detected in the given sample with fixed sample size. Alternatively in a post hoc analysis of the test result after obtaining output of the proposed system, one can determine the minimum effect size this comparison task can achieve at a given power level. Practically in analyzing an already completed comparison task, researchers can specify a desired power level, significance level and the sample size in order to determine the effect size of the planned comparison task. In this thesis, however, we will use the power estimation methods to estimate statistical power for candidate tests in the case studies in a post hoc fashion, where we attempt to understand how the sample size is related to power.

2.8 Hypothesis testing tasks

To evaluate and compare system performance, we would like to design a hypothesis testing procedure specific to NLP systems. At first glance, one tends to directly compare the magnitude of values of the chosen evaluation metric of the two competing systems, i.e., $\mathbf{X} > \mathbf{Y}$

or $\mathbf{X} < \mathbf{Y}$ or $\mathbf{X} = \mathbf{Y}$. Knowing the exact distributions of $\mathbf{X}$ and $\mathbf{Y}$ and using basic probability theory, one could of course derive the probabilities of the first and second inequalities, whereas the third equality will also give a probability of 0 if $\mathbf{X}$ and $\mathbf{Y}$ follow continuous distributions. However, it may not be meaningful to directly compare two random variables; rather, from the frequentist point of view, since the parameters which determine the distribution function of a random variable are unknown constant, it may interest researchers to compare the parameters instead, for example, the population mean. Hence, adopting this idea, the appropriate conclusion to draw must be drawn with respect to the population mean.

Now, by denoting different sets of hypotheses to be tested as $Task$, we introduce several exemplary tasks which take different null hypotheses. One is to test for the equality of means and another is to test for identical distribution. In this section and forward, we will use regular statistical notations instead of notations specific to testing in NLP. It is important to note that the hypothesis testing described here follows the frequentist approach, whereas hypothesis testing can be accessed from a Bayesian point of view as well, which will not be discussed in this thesis.

Henceforth, we will denote a hypothesis testing task by **T1, T2, T3, ...**

2.8.1 Testing location difference

Many probability distributions are parametrized by a location parameter, such as the mean in the normal distribution and the location in the Cauchy distribution. Researchers intend to make inference about the samples with respect to location shift or a central tendency of a distribution in a two-sample testing problem–that is how the two samples differ on average. Under this consideration, the following hypothesis testing task is given.

Task 1 (**Equal mean).** Consider a two-sample testing problem where we are interested to investigate the equality of means. Suppose $X_1, \cdots, X_n \sim F_X$ and $Y_1, \cdots, Y_m \sim F_Y$, where F_X and F_Y denote the probabilistic distributions from which X_i and Y_j are sampled.

Denote the means of X_i and Y_j by $\mu_X = \mathbb{E}X_i$ and $\mu_Y = \mathbb{E}Y_j$. We consider the following hypothesis testing problem.

$$H_0 : \mu_X \leq \mu_Y \quad v.s. \quad H_1 : \mu_X > \mu_Y \tag{2.21}$$

The null hypothesis is that on average X_i is less or equal to Y_i; the alternative states that on average X_i is greater than Y_i. Note that this two-sample testing can be rewritten as a one-sample testing problem by only considering the true mean difference $\mu_X - \mu_Y = \mu_Z$ being smaller or equal to 0 as the null hypothesis, if $n = m$, where $Z_i = X_i - Y_i$.

$$H_0 : \mu_Z \leq 0 \quad v.s. \quad H_1 : \mu_Z > 0 \tag{2.22}$$

Task 1 is commonly used and appears appropriate when average performance comparison is of interest and the given data is numerical and cardinal. Note that it is not necessary to have '$\leq$' sign in the null hypothesis; one could choose to phrase the null and alternative hypotheses as either a two-sided or a one-sided problem. A two-sided hypothesis has a null that the population means are equal; the alternative is that the two population means are not equal. A one-sided hypothesis is similar to Task 1.

Note that in NLP system performance comparison, we are mostly facing the matched pair design, meaning the two samples are paired. In this case, we could reduce the two-sample testing problem to a one-sample testing problem by simply considering the difference of the two samples. The corresponding null hypothesis is that the mean of the difference is 0, and the alternative is the negation of the null.

2.8.2 Testing identical distribution

Without knowledge of the exact probability distributions of the given samples, one may alternatively ask how their distributions differ from each other–that is whether or not the two samples are from the same distribution. Such null hypothesis arises frequently in NLP system comparison tasks, where people intend to assume in their null hypothesis that the two systems produce similar results–that is the system outputs of the two systems display no statistically significant difference.

Task 2 (**Identical distribution).** *Consider the following hypothesis testing problem.*

$$H_0 : X \sim Y \quad v.s. \quad H_1 : X \overset{sto.}{\geq} Y \tag{2.23}$$

where $\sim$ *denotes that* X *is identically distributed as* Y*–that is* X, Y *are from the same distribution, and* $\overset{sto.}{\geq}$ *denotes that* X **stochastically dominates** Y*, as described in Definition 15.*

Lemma 1. *If* A *stochastically dominates* B*,* $\mathbb{E}_A(x) > \mathbb{E}_B(x)\ \forall x \in D$*.*

Therefore, by Lemma 1, we have

$$X \overset{sto.}{\geq} Y \implies \mu_X > \mu_Y \tag{2.24}$$

which suggests that accepting the alternative hypothesis implies the expectation of F_X *is strictly greater than the expectation of* F_Y*, or in plain words,* F_X *is larger than* F_Y *on average, which is equivalent to the alternative hypothesis in the hypothesis testing Task 1. Trivially, the null hypothesis also implies that* $\mu_X = \mu_Y$*. Therefore, hypothesis testing Task 2 implies hypothesis testing Task 1.*

2.8.3 Testing independence

Some may be interested in knowing whether or not two random variables are statistically independent. For example, given two categorical samples, one is interested in understanding the association of the two samples with respect to their categories. Then the following testing hypotheses are appropriate.

Task 3 (**Independence).** One could also be interested in knowing whether two samples are independent or not. As usual consider two samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$, of the same sample size, which are observed in pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$.

We consider the following hypothesis testing problem.

$$H_0 : X \perp\!\!\!\perp Y \quad v.s. \quad H_1 : X \not\perp\!\!\!\perp Y \tag{2.25}$$

The samples X and Y can be numerical or categorical.

2.8.4 Testing general parameters

Alternatively, if the distribution of values of the chosen evaluation metric is asymmetric, one would tend to choose to compare the medians of the two distributions. The set of hypothesis tests that are premised on a parameter of a proposed model or of a distribution or some order statistic (quartiles or median) gives rise to a general testing hypothesis. Also, testing whether or not the Pearson correlation coefficient is 0 (no association) is under this set of general hypotheses. Note that the hypotheses need not be phrased in the following way, but to the user's discretion and choice.

Task 4 (Other parameters). One could also consider a test based on other quantities such as the median or a specific parameter (parameters) of the probabilistic model assumed on the data. Denote the parameters of interest by θ_X and θ_Y. Consider the following hypothesis testing problem.

$$H_0 : \theta_X \leq \theta_Y \quad v.s. \quad \theta_X > \theta_Y \tag{2.26}$$

Common choices of test statistics θ are the Pearson correlation coefficient, the proportion or probability p in a binomial distribution and medians. Testing for median equality is a more robust alternative to testing for mean equality when the underlying distribution is skewed and asymmetric or when outliers are present in the data.

In nonparametric tests such as the sign test, one usually would like to test whether the distribution of the given sample is symmetric around a specific number or not. In this case, under the null hypothesis that the distribution is symmetric around 0 for example, it is equivalent to testing the median being 0 in T4. Note that under Task 2, if the null hypothesis cannot be rejected, it might imply that there is no statistically significant evidence against the null and thus the two samples may come from the same distribution, meaning that they have a common mean, variance or other higher moments, if meaningfully defined, as well. If the null hypothesis is rejected by a test, then we can only say that the two samples are significantly different in their distributions, which does not necessarily suggest the existence of different moments such as the mean and the median.

2.9 Table of tests

The following is a simplified table of tests to be introduced in Chapter 4. Note that all tests assume random sampling defined in Assumption 6. Assumptions A1 to A4 are assumed and will not be tested and included in this table.

Table 2.4: The table of tests (simplified)

Test name	Parametricity	Assumptions	Task	Section
t test	✓	A5, A7, A8, A9, A10	T1	4.2.1
paired t test	✓	A7, A8, A9, A10	T1	4.2.2
F test	✓	A5, A7, A8, A9	T4	4.1.3
χ^2 GOF test	×	A5, A7, A8	T2	4.3.1
χ^2 independence test	×	A7, A8	T3	4.3.2
McNemar's test	×	A7	T4	4.3.3
Wilcoxon signed-rank test	×	A7	T4	4.3.5
Mann-Whitney U test	×	A5, A7	T4	4.3.6
Sign test	×	A7	T4	4.3.8

Chapter 3
PREVIOUS WORK

This chapter will survey numerous research papers in the NLP fields from two perspectives. We will start with methodological papers which focus on introducing, comparing and proposing statistical tests for NLP system performance comparison, where we will direct our attention to their setting of hypothesis testing, assumptions of the tests and conclusions drawn from the test results. Secondly, in empirical work where authors employ statistical tests to perform system performance comparison, we will concentrate on how the authors use the tests, what assumptions are made, how the satisfaction of the assumptions are checked and what conclusions are drawn.

3.1 Methodological work

As numerous newly-invented NLP algorithms emerge, it is necessary to make valid comparisons among competing algorithms with respect to how they perform given some test set, where researchers usually resort to statistical significance testing methods. Despite accentuated precedence and extensive use of statistical testing in this field, it has been recognized that some practices appear invalid and even erroneous. Dror et al. (2018) mentioned in their guide to testing statistical significance in NLP that out of 180 experimental long papers accepted by ACL 2017, 63 of them checked for statistical significance of their proposed systems; 42 of them mentioned the concept of statistical testing, and only 36 papers used the correct statistical tests. The severity of the status-quo could exacerbate in face of the fact that since a multitude of researches relies on multiple comparison testing, the adjustment to multiplicity to control for family-wise significance level is overlooked.

Dror et al. (2018) claims that statistical significance testing is often ignored or misused in

Table 3.1: The table of methodological work.

Work	**Tests**	**Tasks**
Dror et al. (2018)	paired t test, sign test, McNemar's test, Wilcoxon signed-rank test, Pitman's permutation test, paired bootstrap test	T1, T4
Demšar (2006)	paired t test, Wilcoxon signed-rank test, sign test, ANOVA, Friedman test	T1, T4 (median)
Yeh (2000)	χ^2 test, paired t test, sign test, Wilcoxon, randomized paired t test	T1, T2, T4 (median)
Clark et al. (2011)	stratified approximate randomization test	T3
Smucker, Allan, and Carterette (2007)	randomization test, Wilcoxon, sign test, bootstrap, t test	T1, T3

NLP literature. Thus, they intend to introduce useful statistical tests and a simple protocol for choosing the correct statistical test. This paper starts with building the theoretical framework for statistical hypothesis testing. Most importantly, various measurements and metrics considered in the NLP field are mentioned and are deemed to be crucial to the statistical testing procedure as they greatly impact what test one should choose. It should be noted that in this paper, a rather simplified testing problem is proposed, instead of making comparison on two algorithms using multiple test sets as in the replicability study. Then, it introduces parametric and non-parametric significance tests and the test selection process. In the conclusion, this paper mentions current issues with NLP researches and discuss open questions on the influence of independence of data and cross validation to statistical testing.

Demšar (2006) compared several statistical significance tests specific to classifier performance comparison using multiple data sets. This work surveyed accepted papers of the International Conference on Machine Learning from 1999 to 2003 with respect to what tests and evaluation metrics were used to conduct system comparison. The author compared the paired t test, the Wilcoxon signed-rank test and the sign test in a two-system comparison setting; also, ANOVA (analysis of variance) and the Friedman test were introduced in multiple system comparison. Most importantly, the author compared the statistical tests from different angles: statistical assumptions, power and replicability. This work concluded that even though researches may be inclined to the notion that comparing systems across different data sets could result in a more generalized conclusion, there exists no standard for making such comparison, and also tests employed in this case usually are premised upon dubious statistical foundations, which may produce unverified results. This work also recommended the use of nonparametric tests.

Yeh (2000) also examined applicable statistical tests in regards to their different assumptions on the data. This work is primarily concerned with addressing the independence assumption which is said to be violated often in NLP experiments. Yeh suggested that in testing for simple metrics such as recall, matched pair t test and the Wilcoxon test can be applied, whereas more complicated metrics such as precision and balanced F-scores are chosen as the evaluation metric, randomization methods may be more appropriate. This work also pointed out dependencies between test sets may pose a problem in significance testing for system performance.

Clark et al. (2011) took the effect of the optimizer instability variable into account and employed the stratified approximate randomization test to compare system performance across test sets. They also recommended that replication of empirical results be adopted as standard practice.

Smucker, Allan, and Carterette (2007) focused on comparing different statistical significance tests in information retrieval evaluation. They used randomization tests, the Wilcoxon signed rank test, the sign test, the bootstrap test and the student t test as their methods

to compute p-values, from the outputs of TREC 3 (Text REtrieval Conference). Their work concluded that the randomization test, the bootstrap shift method and the student t test produce comparable p-values with no practical difference.

In their descriptions of the tests, the authors noted the relevant set of hypotheses to be tested. The randomization test relies on the null hypothesis that two systems are identical in their distributions. The Wilcoxon signed-rank test and the sign test are for the same null hypothesis as the randomization tests. The bootstrap test's null hypothesis is that the scores of two systems are random samples from the same distribution, which is different from the null hypotheses of previous tests. The student t test takes the null hypothesis that two systems produce scores which are distributed according to the same normal distribution.

Notably, they claimed that the Wilcoxon signed rank test and the sign test produce different p-values, and may incorrectly predict significance and even fail to identify significant results, for which they advised discontinuation of use of these two tests. However, this work did not explicitly mention statistical assumptions made by each test or examine whether the statistical assumptions are met in their example.

3.2 Empirical work

This section will discuss previous work which used statistical tests as tools to compare system performance with respect to the questions the authors attempted to address and presumptions made. Here, we only list an incomprehensive selection of example studies to demonstrate the use of the statistical significance testing in NLP research.

Koehn (2004) conducted a study on the statistical significance tests specifically for machine translation evaluation. This paper presented a well-formed experimental design for system comparison in machine translation, starting with discussing the properties of the evaluation metric BLEU score and selecting the test set. The bootstrap method was used in this work to compute the statistical significance of test results. The author also advocated reporting significance levels in future work for comparing system performance.

Berg-Kirkpatrick, Burkett, and Klein (2012) empirically investigated the practical re-

Table 3.2: The table of example empirical work.

Work	**Tests**	**Tasks**
Berg-Kirkpatrick, Burkett, and Klein (2012)	paired bootstrap	System A is no better than B
Bisani and Ney (2004)	bootstrap	confidence interval
Koehn (2004)	bootstrap	significance of test results
Collins, Koehn, and Kučerová (2005)	sign test	consistent pairwise difference

lation between the statistical significance level and the magnitude of metric gain, the size of the test set and similarity between systems. Also, their work attempted to reveal how well the *i.i.d.* assumption holds in practice. Their paper concluded that the size of the test set constitutes the largest portion of contribution in determining discriminatory ability. Further, given that the test set contains artificially *i.i.d.* observations drawn from the same distribution, then they claimed that significance levels are well-calibrated. Their paper again concurred on the notion that formal testing has its limitations and extra caution should be used when researchers are utilizing statistical tests as the tests provide less information if the target corpus shifts its domain.

Bisani and Ney (2004) used the bootstrap to calibrate the estimation of confidence intervals in Automatic Speech Recognition system evaluation tasks. Their work proposed a new analysis of existing evaluation metric of ASR system performance, but the analysis did not include any element of hypothesis testing at all. To employ the bootstrap, they assumed that the test set can be divided into several segments and each segment is independent from the others, which forms the assumption of the bootstrap.

Collins, Koehn, and Kučerová (2005) discussed the use of the sign test based on the bino-

mial distribution described in Lehmann and Romano (2005) (p. 135) and the use of $BLEU$ scores in evaluating machine translation systems. The authors claimed that the sentence-level $BLEU$ score does not give a good measure of translation quality. They defined the per sentence $BLEU$ score alternatively on the entire set of test instances, which necessarily brings correlation within the test set.

Chan, Ng, and Chiang (2007) also applied the sign test and the bootstrap test to evaluate their proposed machine translation system, which is a state-of-art system integrated with word sense disambiguation technique. They concluded that the improvement over the baseline system, although very small, is statistically significant with a p-value less than 0.05, reported by both the sign test and the bootstrap test.

Rush et al. (2012) used the sign test to test significance of their results. In the table of the test results, four different sample sizes and five different languages are given, and the comparison is done between a baseline system and two proposed systems. The authors declared that all results are significant with p-values less than 0.05.

Chapter 4

STATISTICAL HYPOTHESIS TESTING

This chapter will be devoted to introducing statistical significance tests that are relevant to NLP system performance comparison. We will start with tests and approaches which are used to verify the assumptions of statistical significance tests given later in this chapter. Then, we will discuss the classic student t test in both paired and unpaired settings. Then, we will move on to nonparametric tests, such as the Pearson chi-square tests for independence and goodness of fit, the McNemar's test, and the Wilcoxon signed-rank test etc. In each test, we will first discuss its statistical assumptions necessary to perform the test and then the testing procedure including the test statistic. We will also examine the usage of the test in NLP system performance comparison.

In this section, we will describe each test in several aspects: assumptions made by the test, the testing procedure and its usage in NLP system performance comparison task. We will either give a general testing procedure or a specialized testing example for a clearer illustration. We will also note both potential usage and existing usage.

4.1 Verifying test assumptions

It is necessary to verify whether the assumptions made by the test are satisfied before conducting a certain test. This section will explore tests and approaches to verify test assumptions.

Table 4.1 outlines a few verifiable assumptions and their corresponding verifying tests or methods. In verifying the normality assumption, the usually used method is the Shapiro-Wilk test. There are also other tests such as the Cramer-von-Mises test and the Pearson's test for normality. These tests are implemented in R and *Python*. The symbol '/' denotes

Table 4.1: The table of tests to verify assumptions ('/' denotes the test is not covered in this thesis)

Assumption	**Verifying methods**	**Method assumption**
Independent samples (A5)	Pearson chi-square test for independence, permutation test	A7, A8
Identical distribution (A8)	K-S test, Anderson-Darling test	/
Normality (A9)	Shapiro-Wilk test	A7, A8
Homoscedasticity (A10)	F test	A5, A7, A8, A9
Linearity (A12)	graphically, linear regression	No assumption
Monotonicity (A13)	graphically	No assumption

either that no assumption is made or that the corresponding tests are not introduced in this thesis.

In verifying the assumption of identical distribution, the Kolmogorov-Smirnov (K-S) test and the Anderson-Darling test are both frequently used. We will not cover the two tests in this thesis.

In verifying the assumptions of linear association and monotone association, the usual way is by graphical inspection, such as a scatter plot of the two samples or a simple linear regression line. If the two samples are a priori known to be ordinal, then by definition the monotone association is implied.

Again, Assumptions A1 to A4 will not be tested but rather assumed.

4.1.1 Assumptions on the samples

Assumptions on the samples such as the *i.i.d.* assumption (A7, A8), independent samples (A5), exchangeability (A6) and representativity (A11) cannot be verified directly; they serve as the criteria, according to which the test data are collected. For example, if a statistical test requires *i.i.d.* samples, then each observation in the sample must be collected independently; the identical distribution assumption on each observation within each sample is taken for granted since there is no way to verify it.

4.1.2 Normality

The normality assumption (A9) can be verified by many tests, which are usually called the normality tests.

Central limit theorem

The Central limit theorem (CLT) is a classic statistical theorem which ensures the asymptotic normality of sample means of *i.i.d.* samples. If the evaluation metric is the sample mean of *i.i.d.* samples, then by the CLT, as the sample size increases to infinity, asymptotically the evaluation metric is normally distributed.

Shapiro-Wilk normality test

The Shapiro-Wilk normality test (Shapiro and Wilk, 1965) is a frequently used test for normality. 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

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\left(\sum_{i=1}^{n} X_i - \bar{X}\right)^2} \tag{4.1}$$

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

$$a' = (a_1, \cdots, a_n) = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{1/2}} \tag{4.2}$$

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. The classic way to obtain the p-value in this test is to look up in a table. However, this test is implemented in R and *Python* and can be easily accessed.

4.1.3 Homoscedasticity

The assumption of homoscedasticity described in A10 can be verified by the F test for comparing variances (Snedecor, 1934). Consider two *i.i.d.* samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ which each follow a normal distribution.

Assumptions

1. $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ are independent samples (A5).

2. X_i's and Y_j's are *i.i.d.* and normally distributed (A7, A8, A9).

Testing

Consider the following hypothesis testing problem

$$H_0 : S_X^2 \leq S_Y^2 \quad v.s. \quad H_1 : S_X^2 > S_X^2 \tag{4.3}$$

We have the following testing procedure outlined in Table 4.2.

Note that the F test relies on normality assumption and empirically displays high sensitivity to non-normality. Thus, before conducting the F test, one should examine if normality is satisfied first. Alternatively, one could resort to graphical investigation of the spread of the histograms of the two samples.

4.1.4 Linear association

The linear association assumption (A12) can be easily verified by graphical inspection. Plot two samples against each other and see if there exists a linear relationship. Alternatively,

Table 4.2: Testing procedure for F test

- Given the samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$, the expected values of the two samples can differ. We compute the sample averages $\bar{X}$ and $\bar{Y}$.

- Compute the sample variances S_X^2 and S_Y^2.

- Compute the observed F statistic
$$F = \frac{S_X^2}{S_Y^2} \tag{4.4}$$

- The observed F statistic follows an F distribution with $n-1$ and $m-1$ degrees of freedom. Let $f_{n-1,m-1}$ be a random variable following an F distribution with $n-1, m-1$ degrees of freedom. Then, the p-value is given by
$$p\text{−value} = 1 - P(F_{n-1,m-1} \leq F) \tag{4.5}$$

- Given a significance level α, if p-value is less than $\alpha \in (0, 1)$, 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.

one could use linear regression to investigate the existence of linear relationship, where a significance test on the coefficients can be done as well.

4.1.5 Monotone association

The monotone association assumption (A13) can also be verified by plotting the two samples against each other. If the resulting plot resembles a monotonic function (monotonically increasing or decreasing), then this assumption is satisfied. Again, note that this assumption is usually checked when the data is ordinal, meaning that the data fall into different categories but each category is meaningful in its value.

4.2 Parametric tests

Statistical significance tests can be roughly categorized into two types according to parametricity. This section will introduce **parametric tests**, which put distributional assumption on the samples, which means that the sample is assumed to follow a specific probabilistic distribution which is known. If such assumption is satisfied, usually parametric tests display greater statistical power than nonparametric tests for the same set of hypotheses.

4.2.1 Unpaired student t test

The (unpaired) student t test is a classic parametric test for testing the equality of means of numerical sample (Student, 1908). The test is used for Task 1. We consider two samples $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{Y} = (Y_1, \cdots, Y_m)$ for the test.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are independent (A5).

2. $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed (A9).

3. Each sample consists of *i.i.d.* observations (A7, A8).

4. Two samples have the same variance (A10).

Testing

Given the hypotheses in Task 1, we have the following testing procedure outlined in Table 4.3.

Table 4.3: Testing procedure for unpaired t test

- Given the samples $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$, compute the sample averages $\bar{X}_n$ and $\bar{Y}_m$, and the sample variances S_X^2 and S_Y^2 and the pooled sample standard deviation

$$S_p = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}} \tag{4.6}$$

- Compute the observed t statistic

$$t = \frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{4.7}$$

- The observed t statistic under the null hypothesis follows a t distribution with $n+m-2$ degrees of freedom. Let T_{n+m-2} be a random variable following a t distribution with $n+m-2$ degrees of freedom. Then, the one-tailed p value is given by

$$p-\text{value} = 1 - P(T_{n+m-2} \leq |t|) \tag{4.8}$$

- Given a significance level $\alpha \in (0, 1)$, if p-value is less than α, we can conclude that there is statistically significant evidence strong enough to reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

Should the assumption of homoscedasticity be violated, there exists a variant of the

student t test which modifies the test statistic and the degrees of freedom of its distribution under the null, called the Welch's t test. One can choose to directly apply the Welch's t test without a priori examination of equality of variances. The modified test statistic is given by

$$t_{welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_X^2}{m}}} \tag{4.9}$$

which under the null follows a t distribution with ν degrees of freedom, where

$$\nu \approx \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}} \tag{4.10}$$

Usage in NLP

The unpaired version of the student t test is testing for mean difference and takes on T1 as its hypothesis. If a system performance comparison task focuses on a conclusion based on the mean (average performance), and the data conform to the test assumptions aforementioned, then the student t test is appropriate and would even display more power than other nonparametric tests. However, note that the unpaired t test assumes independent samples, but many performance comparison tasks are matched pair designs, which implies its limited practical usage in NLP.

Since the t test is a parametric test, it is important to verify the assumptions before conducting the test. However, the normality assumption can be relaxed given large sample size ($\gg 100$), which arises from an empirical conclusion that the t test is not sensitive to the absence of normality, compared to the F test. Nevertheless, the assumption that the two samples contain *i.i.d.* observations must be satisfied in order to properly use the test. This assumption cannot be verified but rather is determined in the process of data gathering or sampling. If the sample comes from a random sampling procedure, then the sample contains independent observations; the identical distribution assumption then must be taken for granted.

The data type required by the t test is numerical. In NLP, for example, accuracy, $BLEU$ score, $ROUGE$ score and other evaluation metrics of which numerical values denote certain meanings are the suitable data type. However, note that even though the normality assumption can be relaxed, the normal distribution ranges over the entire real number $\mathbb{R}$, and metrics such as accuracy, $BLEU$ and $ROUGE$ are all real numbers ranging from 0 to 1. In this sense, the t test is not applicable.

To give an example of applicable case where the t test can be used, students from two classes instructed by different teachers took a certain test. The school is interested to know how differently the two classes of students performed on the test. Given that the test scores are normally distributed and each student took the test independently from other students, the unpaired student t test can be used to detect mean difference of test scores between the two classes of students.

4.2.2 Paired student t test

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 assumption is similar to that of the unpaired version but requires that $\mathbf{X}$ and $\mathbf{Y}$ are dependent.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired and thus dependent.

2. $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed (A9).

3. Each sample consists of *i.i.d.* observations (A7, A8).

4. Two samples have the same variance (A10).

5. Sample sizes of $\mathbf{X}$ and $\mathbf{Y}$ must be the same.

Testing

Given the hypotheses in Task 1, 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

$$t_{paired} = \frac{\bar{Z}}{S_Z/\sqrt{n}} \tag{4.11}$$

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.

Usage in NLP

The paired version of the student t test possesses more applicability than the unpaired version in NLP system performance comparison due to its matched pair nature (Sadat, Yoshikawa, and Uemura, 2003; Chen, Mukherjee, and Liu, 2014; Jang et al., 2016). However, it is still a parametric test replying on heavy assumptions on the data. It is necessary to verify whether or not these assumptions are satisfied before conducting the t test.

4.3 Nonparametric test

This section presents several **nonparametric** statistical significance tests, which do not put any distributional assumptions on the data, as opposite to parametric tests. Nonparametric tests usually have fewer and looser assumptions on the data than parametric tests do and thus can be applied to a wider range of hypotheses, which however necessarily decreases the statistical power nonparametric tests can have. If distributional assumptions on the data can be verified, then a parametric test is preferable; otherwise, a nonparametric test may be more appropriate.

4.3.1 Pearson chi-square test for goodness of fit

The Pearson chi-square test is a nonparametric test for categorical data, which can be used in two different ways: the first use of the Pearson chi-square test is for goodness of fit

test (testing if the sample follows a specific distribution) and the second use is to test for independence of two categorical random variables (Pearson, 1900).

The Pearson chi-square test can be applied to one-sample problems and two-sample (contingency table) problems as well. In a one-sample problem with categorical data, one is usually interested to test whether the given data is compatible with a proposed probabilistic model or not. For example, to test whether a six-faced die is fair or not, we can use the Pearson chi-square test of goodness of fit to investigate the divergence of probabilities of getting each number from the expected fair probabilities. In a two-sample problem, one is usually interested in knowing if the distributions of the two samples are the same, in which case an expected distribution is not available and thus estimating the expected probabilities becomes necessary. The following two examples illustrate the use of the Pearson chi-square test for goodness of fit and for independence of two categorical variables.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are independent (A5).

2. Each sample consists of *i.i.d.* observations (A7, A8).

3. Expected count in each cells exceeds 5.

Testing

Consider rolling two six-faced dice independently both for n times. We then have $X_1, \cdots, X_n \in \{1, \cdots, 6\}$ and $Y_1, \cdots, Y_n \in \{1, \cdots, 6\}$. We are interested in knowing if the distributions of the two dice are the same. Thus the null hypothesis is $X \sim Y$ and the alternative is the simple negation. Define observed counts $N_X(s) = \#\{i : X_i = s\},\ N_Y(s) = \#\{i : Y_i = s\}$ for $s \in \{1, \cdots, 6\}$. Under the null hypothesis, $X \sim Y$. The expected counts are given by np_s where $p_s = P\big(X_i = s\big) = P\big(Y_i = s\big)$, for $s \in \{1, \cdots, 6\}$. However, we do not know p_s; it is

necessary to estimate p_s using the combined sample of X and Y by

$$\hat{p}_s = \frac{N_X(s) + N_Y(s)}{2n} \tag{4.12}$$

for $s \in \{1, \cdots, 6\}$. Thus the expected counts become $n\hat{p}_s$. The test statistic is given by

$$D = \sum_{s=1}^{6} \left[\frac{(N_X(s) - n\hat{p}_s)^2}{n\hat{p}_s} + \frac{(N_Y(s) - n\hat{p}_s)^2}{n\hat{p}_s} \right] \tag{4.13}$$

Under the null, the test statistic D asymptotically follows χ^2_d. The degree of freedom d is given by $(m-1)(S-1)$ where m is the number of dice and S is the number of faces. In this case, $d = (2-1)(6-1) = 5$. Note that in this example, the number of rolls for each die can be different.

Usage in NLP

The Pearson chi-square test for goodness of fit can be applied to classification tasks with dichotomous results, for example a classifier that outputs a result of either true or false (Graham, Mathur, and Baldwin, 2014; Ghaeini et al., 2018). In the work by Graham, Mathur, and Baldwin (2014), they used the Pearson chi-square test to determine whether or not the combination of metric and test is accurate. Ghaeini et al. (2018) employed the Pearson chi-square test to assess the performance improvement of the proposed natural language inference technique from a baseline system.

In NLP tasks such as POS tagging, a contingency table of correct and wrong predictions can also be formed, which then can be subjected to the Pearson chi-square test for goodness of fit. Note that the null hypothesis in this case is phrased with respect to the distribution of values of the chosen evaluation metric–that is the two distributions are the same or in plain words, two systems display no difference in the results (T2). In addition, the Pearson chi-square test can be potentially used to determine whether a sample is from a particular distribution by calculating the Pearson chi-square statistic in the histogram of the sample compared with that of the proposed distribution. The test can also be used as a normality test with relatively low power.

The Pearson chi-square test for goodness of fit assumes independence between the two systems and *i.i.d.* observations in each sample, two assumptions which often appear to be strong, since usually we are facing a matched pair design. Therefore, the Pearson chi-square test for goodness of fit is not recommended for comparing two systems.

4.3.2 Pearson chi-square test for independence

In a similar example aforementioned where we throw the two dice together, we would like to know if the two dice are independent.

Assumptions

1. Each sample consists of *i.i.d.* observations (A7, A8).

2. Expected counts in all cells exceed 5.

3. Categorical data is required.

Testing

Thus the null hypothesis is that $X \perp\!\!\!\perp Y$ and the alternative is the simple negation. Assuming that each throw is independent of each other, we obtain a sample of pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$. Under the null, the two dice are independent; thus $\forall s, t \in \{1, \cdots, 6\}$, by the definition of independence, the joint probability equals the product of the marginal probabilities, i.e.,

$$P\big(X = s, Y = t\big) = P\big(X = s\big)P\big(Y = t\big) \tag{4.14}$$

However, we do not know the marginal probabilities, and it is necessary to estimate them from the data. Define marginal counts similarly by $N_{s+} = N_X(s)$ and $N_{+t} = N_Y(t)$. The estimated marginal probabilities are

$$\hat{p}_s = \frac{N_{s+}}{n} \qquad \hat{q}_t = \frac{N_{+t}}{n} \tag{4.15}$$

Thus, $\forall s, t \in \{1, \cdots, 6\}$, the estimated expected counts are given by

$$n\hat{p}_s\hat{q}_t = \frac{N_{s+}N_{+t}}{n} \tag{4.16}$$

The test statistic is given by

$$D = \sum_{s=1}^{6}\sum_{t=1}^{6} \frac{(N_{s,t} - n\hat{p}_s\hat{q}_t)^2}{n\hat{p}_s\hat{q}_t} \tag{4.17}$$

Under the null hypothesis, the test statistic D follows a χ^2 distribution with $(S-1)(T-1)$ degrees of freedom where S and T are the number of faces of the two dice. In this case, the degrees of freedom are $(6-1)(6-1) = 25$.

Usage in NLP

The Pearson chi-square test for independence is used to test for independence between two categorical variables (T3). In NLP, such demand rarely arises in comparing system performance, except for the case where one is interested in knowing the association between two systems which produce categorical outputs to draw a conclusion that the two systems are independent in their system outputs. However, this case rarely is needed, since it is obvious that given two good POS taggers, for a particular token correctly predicted, the association will be positive. Trinh, Ross, and Kelleher (2019) used the Pearson chi-square test for independence to determine the dependencies between dialogue state slot variables and the extent to which the dependencies are present. Štefanec, Ljubešić, and Kraljević (2016) performed a chi-square test to verify whether the number of correctly and incorrectly spelled tokens and the clinical status of the subject are independent.

4.3.3 McNemar's test

McNemar's test is similar to the Pearson chi-square test but it is for data of matched pairs (McNemar, 1947). The matched-pair design of experiment usually occurs in comparison of two treatments on the same subject, which yields a pair of two responses. In essence, the

test is comparing the proportions of responses that exhibit disagreement. In Table 4.4, we can note that testing for equality of p_{21} and p_{12} implies that $p_{1+} = p_{+1}$ and $p_{2+} = p_{+2}$, where p_{i+} denotes the proportion of row i and p_{+j} denotes the proportion of column j, resulting in testing for equality of marginal proportions. This testing scenario falls into Task T4 specialized for marginal proportions. Now, note again that the two samples are from binomial distributions. A binomial distribution has two parameters: the probability p and the sample size N. Given the same sample size N, if the probabilities p agrees, then the two samples have the same distribution. Thus, testing for T4 is equivalent to testing for T2.

Assumptions

1. Two samples both contain independent observations (A7).

2. Two groups, $\mathbf{X}$ and $\mathbf{Y}$ and dependent nominal (or categorical) variable.

3. Observations in both groups can only submit to one category.

Testing

Consider two categorical random variables $\mathbf{X} \in \mathbb{D}_X$ and $\mathbf{Y} \in \mathbb{D}_X$, which form a contingency table of dimension $\dim(\mathbb{D}_X) \times \dim(\mathbb{D}_Y)$, where $\mathbb{D}_X$ and $\mathbb{D}_Y$ are the sample spaces. For simplicity assume that $\mathbf{X}$ and $\mathbf{Y}$ both have two categories, denoted by $\{1, 2\}$. Define cell counts and probabilities by $N_{st} = \#\{X = s, Y = t\}$ and $p_{st} = P\big(X = s, Y = t\big)$, $\forall s, t \in \{1, 2\}$.

Table 4.4: The contingency table of samples $\mathbf{X}$ and $\mathbf{Y}$

	1	2
1	N_{11}	N_{12}
2	N_{21}	N_{22}

Consider Table 4.4 where the row represents treatment A and the column represents treatment B, and responses are either *success* denoted by 1 or $failure$ denoted by 2. Then the null hypothesis of no treatment effect is $p_{12} = p_{21}$. The test statistic is given by

$$D = \frac{(N_{12} - N_{21})^2}{N_{12} - N_{21}} \tag{4.18}$$

Under the null, given large numbers of N_{12} and N_{21}, the test statistic D follows a χ^2 distribution with 1 degree of freedom.

Usage in NLP

This test is used for results which can be formulated into a 2×2 contingency table (the table of recall and precision) in order to compare disagreement of predictions (Blitzer, McDonald, and Pereira, 2006). Given that the null hypothesis is that there is no difference between the two systems, in NLP system comparison tasks, it is more appropriate to say that the two systems display no difference in misclassifying the test set (Task 4). This test appears to posses more applicability than the Pearson chi-square test for goodness of fit in that it does not assume independence between the two samples.

4.3.4 Fisher's exact test

The Fisher's exact test (Fisher, 1922) is used in order to determine whether two binary categorical variables are independent or not (Fisher, 1922). Traditionally, the Fisher's exact test was invented to resolve the problem in a 2×2 contingency table of insufficient expected cell counts, a condition which prohibits the use of the chi-square test. In addition to the case where the failure of the Pearson chi-square test occurs, the Fisher's exact test assumes fixed row and column counts (not random).

Assumptions

1. Fixed column and row total counts.

2. Observations are independent (A7).

Testing

Consider binary categorical variables $X, Y \in \{1, 2\}$ and a contingency table identical to Table 4.4. Under the null that $X \perp\!\!\!\perp Y$, the probability of obtaining such a table, conditional on having the observed marginal counts, is

$$\binom{N_{1+}}{N_{11}}\binom{N_{2+}}{N_{21}}/\binom{n}{N_{+1}} = \frac{N_{1+}!N_{2+}!N_{+1}!N_{+2}!}{N_{11}!N_{12}!N_{21}!N_{22}!n!} \tag{4.19}$$

where the top left cell count is of a hypergeometric distribution. The p-value is the sum of the probability of the observed table and the probability of the most extreme table, where the marginal counts are fixed and cell counts are permuted.

The Fisher's exact test can be generalized to contingency tables of any sizes.

Usage in NLP

The Fisher's exact test is used when in the Pearson chi-square test the assumption that each cell must contain at least 5 expected counts is not satisfied and when the margins are known to be fixed. One advantage associated with the Fisher's exact test is that the p-value is calculated exactly, in comparison with the approximated p-values of the Pearson chi-square test. In NLP system comparison tasks, the Fisher's exact test is used for testing association between two categorical variables (independence), which corresponds to the Pearson chi-square test for independence (Task 3). Again, need for such tests seldom arises. Also, the assumption of fixed row and column counts appears to be unrealistic.

Notably, the Fisher's exact test is used in NLP for purposes other than system performance comparison. Weeber, Vos, and Baayen (2000) used the Fisher's exact test to compute the significance of common word association for a medical information extraction system when the frequency of side-effect-related terms is less than 5, which violates the assumption of the Pearson chi-square test. Yang and Zheng (2009) used the Fisher's exact test to prune the translation table of a hierarchical phrase-based statistical machine translation system.

4.3.5 Wilcoxon signed-rank test

The Wilcoxon signed-rank test is the nonparametric counterpart to the paired student t test. It attempts to test the hypothesis that the difference between the two samples are symmetric around 0. Since this test is based on ranks, the samples must be ordinal or at least can be ranked (Wilcoxon, 1945).

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired.

2. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are independent (A7).

3. Distributions of $\mathbf{X}$ and $\mathbf{Y}$ are symmetric.

Testing

Let $Z = X - Y$. Consider the following hypothesis testing problem.

$$H_0 : Z \textit{ is symmetric around } 0. \;\; v.s. \;\; H_1 : Z \textit{ is not symmetric around } 0. \tag{4.20}$$

We have the following testing procedure outlined in Table 4.5.

The Wilcoxon signed-rank test is implemented in R and can be easily accessed.

Usage in NLP

The Wilcoxon signed-rank test is a frequently used nonparametric test in comparing NLP system performance (Søgaard, 2013; Søgaard et al., 2014) . It is preferred due to its looser assumptions on the data than parametric tests. Note that the test assumes symmetry of the distributions of the two samples, and it is testing for the location difference of the medians rather than the means, which suggests that the test is not sensitive to the presence of outliers, in comparison with the t test. However, the test still assumes independence between pairs.

Table 4.5: Testing procedure for Wilcoxon signed-rank test

- Calculate $|Z_i|$ and $\text{sign}(Z_i)$, where sign is the sign function.
- Discard Z_i's which are 0, and let n_r be the new sample size.
- Sort $|Z_i|$ in ascending order.
- Rank $|Z_i|$, and let the rank of ties be their average rank.
- Calculate
$$W = \sum_{i=1}^{n_r} \text{sign}(Z_i) \cdot \text{Rank}(|Z_i|) \tag{4.21}$$
- Under the null hypothesis, as $n_r \to \infty$, W converges to a normal distribution.
- A z-score can be calculated based on W and the p-value can be obtained by looking up in a normal distribution table.

In NLP, the Wilcoxon signed-rank test is appropriate for numerical (continuous) metrics such as accuracy, the *BLEU* score and the *ROUGE* score and for hypotheses which involve location differences. Note that by the symmetry assumption, the mean and the median coincide; therefore, the test is both testing for means and medians (Task 1 and Task 4).

4.3.6 Mann-Whitney U test

The Mann-Whitney U test is a nonparametric test for testing whether two independent samples have the same distribution (Mann and Whitney, 1947). Notice that the Wilcoxon signed-rank test is a similar nonparametric test for dependent samples.

Assumptions

1. **X** and **Y** are independent (A5).

2. Two samples are ordinal or at least can be meaningfully ranked.

3. Observations in **X** and **Y** are independent (A7).

Testing

The test rejects for large values of the test statistic

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{X_i > Y_j\} \tag{4.22}$$

where $\mathbb{I}$ is the indicator function, which takes 1 if the condition is true, and 0 otherwise. For large samples, the test statistic U is approximately normal. The p-value can be obtained by calculating the standardized U to derive a z-score, which is then can be looked up in a standard normal table manually. This test is implemented in both *R* and *Python*.

Usage in NLP

The Mann-Whitney U test is similar to the Wilcoxon signed-rank test except it is for independent samples. The test has another name, which is the Wilcoxon rank sum test. The two tests are equivalent to each other. The test also enjoys the advantage of fewer assumptions on the data and its nonparametricity. In NLP system comparison task, matched pair design is often of interest, which suggests dependent samples. Therefore, unless independent samples are under consideration, the Mann-Whitney U test is not preferred. Yang and Mitchell (2017) used the Wilcoxon rank sum test to compare the results of the proposed model and the baseline models, despite the fact that the comparison should be paired. Amidei, Piwek, and Willis (2019) recommended the use of the Mann-Whitney U test in the evaluation for natural language generation systems.

4.3.7 Permutation test for independent samples

This and the next three subsections will focus on significance tests that utilize randomization. The permutation test is a family of nonparametric and computational tests that basically mimic the original sampling process by resampling without replacement.

The permutation test can be applied to a variety of hypothesis testing problems such as the one described in Task 2, where we are interested in whether the two samples come from the same distribution. The basic idea is that we assume the null is true and resample the observations in a fashion identical to the way the original sample is sampled from the population.

Assumptions

1. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are exchangeable (A6).

The permutation test makes a slightly looser assumption on the sample. Rather than assuming the observations from the sample are *i.i.d.*, the permutation test only assumes exchangeability (6).

Testing

Consider two samples $X_1, \cdots, X_n \sim F_X$ and $Y_1, \cdots, Y_m \sim F_Y$. Under the null hypothesis, $F_X \sim F_Y$. Thus we are allowed to pool the two samples together and shuffle the pooled sample. Since by exchangeability all permutations of the pooled sample have the same probability, we can compute a test statistic of interest in each permutation and compare it to the observed test statistic. Based on the rejection criterion, we count the times when we observe a rejection and calculate the p-value accordingly. The test procedure is given in Algorithm B.1.

As usual, given a significance level α, if p-value is less than α, 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.

Usage in NLP

This version of permutation test is not yet discussed in relevant literature in NLP system performance comparison. It intends to compare the distributions of two samples (T2), with no pairing. The only assumption it has is exchangeability, which is implied by random sampling. However, the matched pair design is the usual case we are facing. Hence, this permutation test demonstrates limited use in NLP system performance evaluation.

4.3.8 Sign test calibrated by permutation

The sign test 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 sign test can also be performed using the binomial distribution instead of calibration by permutation, which will not be discussed here.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired.

2. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are independent (A5).

Testing

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 in Algorithm B.2.

As usual, given a significance level α, if p-value is less than α, 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.

Usage in NLP

The sign test based on the binomial distribution is widely used in NLP system performance comparison (Collins, Koehn, and Kučerová, 2005; Chan, Ng, and Chiang, 2007; Rush et al., 2012), where the major goal is to determine if the pairwise difference between the scores of the two systems displays consistency with respect to positive and negative values. The null hypothesis is that the number of positive differences is the same as the number of negative differences. Under this null hypothesis, it is equivalent to test for the median being equal to 0 or the distribution of the difference is symmetric around 0.

The sign test introduced in this thesis is calibrated by permutation, where we choose a test statistic of interest which demonstrates difference in the null and the alternative. Tests calibrated by permutation methods usually have loose assumptions on the data. The sign test assumes dependent samples (matched pairs) and independence among pairs. In NLP system performance comparison, the sign test may be preferred as it is for matched pair design. However, the assumption of independent pairs still sometimes appears unrealistic. In addition, since the sign test calibrated by permutation is a compute-intensive method, with extremely large samples, the test can be computationally expensive.

4.3.9 Bootstrap test for independent samples

The bootstrap test (Efron and Tibshirani, 1993) is another computational approach which relies on the notion that the given sample approximates the original population and on re-sampling 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.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are independent (A5).

2. Observations in $\mathbf{X}$ and $\mathbf{Y}$ are *i.i.d.* (A7, A8).

3. Samples are representative of their populations (A11).

Testing

The test procedure is given in Algorithm B.3.

Note that the p-value is given by

$$\frac{C+1}{B+1} \tag{4.23}$$

where C is the number of extreme bootstrapped test statistics and B is the number of iterations. There has been disagreement in literature about how the p-value is computed in a bootstrap test. Efron and Tibshirani (1993) suggested that the Achieved Significance Level (ASL) is computed without adding 1 both for the numerator and the denominator, while Davison and Hinkley (1997) computed the p-values as given in Equation 4.23. The purpose of the addition is not explicated in Davison and Hinkley (1997). Suppose that without the addition, $C = 0$, which will result in a p-value of 0. It is possible that one of the intensions of having the addition is to avoid having a 0 p-value.

As usual, given a significance level α, if p-value is less than α, 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.

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \tag{4.24}$$

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

$$[\bar{X} - \bar{Y} - \hat{t}_{1-\alpha/2} \cdot \hat{SE}, \bar{X} - \bar{Y} - \hat{t}_{\alpha/2} \cdot \hat{SE}] \tag{4.25}$$

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

$$\hat{SE} = \sqrt{S_X^2/n + S_Y^2/m} \tag{4.26}$$

The corresponding p-value is defined as the largest α 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.

Usage in NLP

The bootstrap methods are intuitive and easy to understand and to implement. Note that the bootstrap assumes representativity of the sample data, and thus the bootstrap is no cure for small sample size. The regular version of the bootstrap test assumes that observations within each sample are random samples from the population (*i.i.d.*). Based on whether the data is paired or not, the bootstrap technique can differ, as to whether it is necessary to resample with replacement from the pooled samples or from pairs.

The nonparametric bootstrap methods do not assumes any distributions associated with the sample (parametric bootstrap is present but will not be considered). The (non-overlapping) block bootstrap takes the dependencies within each sample into consideration. To sum up, the bootstrap methods, like the permutation methods, are preferred compute-intensive methods to calibrate significance testing when the distributions of the samples are unknown or the distributions of the test statistic are theoretically hard to derive. The resampling technique of the bootstrap methods is dependent on the original sampling of the sample from the population.

We do not separately introduce the bootstrap for paired samples in this thesis. With a slight modification, the bootstrap can be used for testing in a paired two-sample setting by simply considering the pairwise difference, which will reduce the original problem into a one-sample testing. Then, instead of resampling the pooled sample, we need only resample the difference repeatedly with replacement and derive a confidence interval accordingly.

The bootstrap method has gained substantial popularity and is widely used in NLP system performance comparison (Berg-Kirkpatrick, Burkett, and Klein, 2012; Bisani and Ney, 2004; Koehn, 2004; Kumar and Byrne, 2004). In NLP system comparison tasks, one may choose to use the bootstrap methods to calibrate a significance test by first of all assuming representativity of the samples. Then, based on the structure of the samples (paired or unpaired) and possible dependencies within samples, a specific type of bootstrap methods is selected.

The most important consideration in using a bootstrap method is the choice of the test statistic. A 'good' test statistic differs in the null hypothesis and the alternative. For example, the mean difference is 0 under the null and nonzero under the alternative, for tests for location difference.

4.3.10 Block bootstrap

The block bootstrap is a variant of the classic bootstrap method. Similar to block permutation, it respects the dependency structure in the original data and is used for resampling of time series data, for example. The block bootstrap introduced here is specialized for a two-sample testing case with paired data. A block is the minimal unit in the sample that is independent from each other. A block is a subset of evaluation units, and evaluation units within the same block are dependent. Recall that an evaluation unit can be one single test instance or a set of test instances. Simply put, since there is a one-to-one correspondence between the values of evaluation metric and the evaluation unit, the sample considered in the significance testing is a set of values of evaluation metric. Now, to allow dependency among evaluation units, we introduce the notion of blocking.

Assumptions

1. $\mathbf{X}$ and $\mathbf{Y}$ are paired.

2. Blocks are independent.

3. Samples are representative of their populations (A11).

Testing

The test procedure is given in Algorithm B.4.

As usual, given a significance level α, if p-value is less than α, 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.

Again, the choice of the test statistic is up to the user's discretion. However, in this case, the t ratio test statistic may not be applicable since the block structure is under consideration.

Usage in NLP

The concept of the block bootstrap is not yet discussed in significance testing in NLP system performance comparison. The block bootstrap differs from the regular version of bootstrap in that it takes the dependencies among observations in the samples into account. For example, in a machine translation system performance comparison task, the test set usually consists of multiple articles excerpted from different news sources. Each sentence in each article is a test instance, based on which a sentence-level $BLEU$ score can be computed. In a particular article, there is potential dependency among sentences that are close to each other. Also, sentences in a particular article are not generated due to random sampling without replacement, which violates the assumption of independence (but still satisfies exchangeability). Under such consideration, the block bootstrap displays great potential in evaluation tasks of which the test sets contain dependent observations.

4.4 Multiple testing

Sometimes performance comparisons among multiple systems are of interest to researchers. Should such situation arise, adjustment of p-values will become necessary.

4.4.1 Family-wise error rate

Consider the following tests where there are a finite number (s) of systems in the performance comparison. Let μ_i denote the population mean of values of the evaluation metric of system i, for $i = 1, \cdots, n$. Suppose we are interested in knowing whether or not there exists difference in the performance of all s systems. We have the null hypothesis

$$H_0 : \ \mu_1 = \cdots = \mu_n \tag{4.27}$$

while the alternative is that at least two systems do not perform equally. In a naive approach, independently conducting pairwise tests results in an increase in the probability of false rejections or Type I error (Lehmann and Romano, 2005). Therefore, in multiple comparison, the rejection criterion is substituted by the probability of one or more false rejections not exceeding a preset significance level α, which is defined as the family-wise error rate (FWER). Essentially, we require that $FWER \leq \alpha$. There are two simple methods of controlling for FWER.

Bonferroni procedure

The Bonferroni procedure rejects a null hypothesis H_i when $p_i \leq \alpha/s$, where p_i is the p-value derived in hypothesis H_i and s is the number of null hypotheses. This procedure controls the FWER at or below the significance level α.

Holm procedure

The Holm's procedure is also used to control the family-wise error rate and is said to be more powerful than the Bonferroni correction. The procedure is outlined in Table 4.6.

Table 4.6: Holm procedure

- Denote the ordered p-values by $p_{(1)} \leq \cdots \leq p_{(s)}$ and corresponding hypotheses $H_{(1)}, \cdots, H_{(s)}$.
- If $p_{(1)} \geq \alpha/s$, accept $H_1, \cdots, H_s$, and stop; if $p_{(1)} < \alpha/s$, reject $H_{(1)}$ and test the remaining $s-1$ hypotheses at level $\alpha/(s-1)$.
- If $p_{(1)} < \alpha/s$ but $p_{(2)} \geq \alpha/s$, accept $H_{(2)}, \cdots, H_{(s)}$ and stop; if $p_{(1)} < \alpha/s$ and $p_{(2)} < \alpha/(s-1)$, reject $H_{(2)}$ as well and test the remaining $s-2$ hypotheses at level $\alpha/(s-2)$.

The Bonferroni and Holm procedures can be easily accessed in *R* and *Python*.

4.4.2 False discovery rate

The false discovery rate (FDR), developed by Benjamin and Hochberg, which is defined as the expected value of the proportion of falsely rejected null hypotheses, is another element to control for in multiple comparison (Benjamini and Hochberg, 1995).

The Benjamin-Hochberg procedure is outlined in Table 4.7.

Table 4.7: Benjamin-Hochberg procedure

- Denote the ordered p-values by $p_{(1)} \leq \cdots \leq p_{(s)}$ and corresponding hypotheses $H_{(1)}, \cdots, H_{(s)}$.
- Let k be the largest i for which
$$p_{(i)} \leq \frac{i}{s}\alpha \tag{4.28}$$
then reject all $H_{(i)}$ for $i = 1, \cdots, k$.

If the tests are independent (independent test statistics), then the Benjamin-Hochberg procedure controls FDR at α. Also, the authors conducted a power analysis of the FDR controlling method, which shows that the power of the FDR controlling procedure is uniformly larger than that of the other methods, where such advantage increases with the number of non-null hypotheses and in s.

4.4.3 Friedman test

The Friedman test is a rank-based nonparametric test for repeated measures of multiple comparison (Friedman, 1937). Suppose that there are s systems for comparison, denoted

by A_s. Denote the values of evaluation metric by X_i^s and sample size by n. We assume a balanced design, where the sample sizes of all systems are the same. Consider the following table, where the rows represent the value of evaluation metric for a single evaluation unit; the columns represent systems.

Table 4.8: Table of repeated measures

	A_1	A_2	A_3	$\cdots$	A_s
1	X_1^1	X_1^2	X_1^3	$\cdots$	X_1^s
2	X_2^1	X_2^2	X_2^3	$\cdots$	X_2^s
3	X_3^1	X_3^2	X_3^3	$\cdots$	X_3^s
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_n^1	X_n^2	X_n^3	$\cdots$	X_n^s

Assumptions

1. The samples $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are all paired.

2. Observations in each sample are independent (A5).

3. The samples are at least ordinal or can be meaningfully ranked.

Testing

The null hypothesis associated with the test is that there is no difference among the performance of the systems. The test compares ranks within each row. Denote R_{ij} as the rank of X of i-th row and s-th column. The sum of ranks for s-th system is given by

$$R_{+j} = \sum_{i=1}^{n} R_{ij} \tag{4.29}$$

The test rejects for large values of

$$G = \frac{12}{ns(n+1)} \sum_{j=1}^{s} \left[R_{+j} - \frac{n(s+1)}{2} \right]^2 \tag{4.30}$$

Under the null hypothesis, G is asymptotically distributed according to the chi-squared distribution with $s-1$ degrees of freedom. As usual, given a significance level α, if p-value is less than α, 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. Once a significant result is detected, a post hoc test can be performed to unearth which pair of systems contribute to the significance, where the Wilcoxon signed-rank test may become a proper choice.

Usage in NLP

The Friedman test is a less frequently mentioned and used significance test for multiple comparison in NLP system comparison tasks. It is favored since it puts fewer assumptions on the data and conforms to the design of repeated measures. However, it still assumes independence among the observations in each sample (rows). In NLP system comparison tasks, should multiple comparison of systems arise and if the test set contains independent observations, the Friedman test is an appropriate choice of significance test. Demšar (2006) discussed the use of the Friedman test for comparing multiple classifiers across datasets.

4.5 Table of tests (comprehensive)

The following table contains all statistical significance tests given in this thesis. The first column is the name of the test. The second column is the type of test, indicating parametricity of the test. The third column lists the statistical assumptions of the test that are numbered and can be referenced back to previous sections. The fourth column is the data type required by the test, which includes categorical or numerical, paired or unpaired (only noted if paired). The fifth column is the set of hypotheses to which the test is applicable, denoted by T. The last column shows the corresponding test statistic.

Generally, we have two samples $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{Y} = (Y_1, \cdots, Y_m)$, both of possibly different sample sizes n and m. Specifically, if interested, $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$ is the difference of the two samples, where it is required that the sample sizes of $\mathbf{X}$ and $\mathbf{Y}$ are the same, denoted by n.

All tests assume random sampling of the data, defined in A6.

In Pearson chi-square test, O refers to the observed counts and E refers to the expected counts.

In McNemar's test, N_{12} and N_{21} are defined in Table 4.4. Also, the McNemar's test assumes that there are large enough counts of N_{12} and N_{21}.

In the sign test, the permutation test and the bootstrap test, the test statistic is determined by the user based on a particular test case.

The symbol '/' implies the test does not have the shared assumptions given in this thesis.

Table 4.9: The table of tests

Test name	**Parametricity**	**Assumptions**	**Data type**	**Task**	**Test statistic**
t test	✓	A5, A7, A8, A9, A10	numerical	T1	$\frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$
paired t test	✓	A5, A7, A8, A9, A10	numerical, paired	T1	$\frac{\bar{Z}}{S_Z/\sqrt{n}}$
F test	✓	A5, A7, A8, A9	numerical	T4	$\frac{S_X^2}{S_Y^2}$
χ^2 GOF test	×	A5, A7, A8[a]	categorical	T2	$\sum \frac{(O-E)^2}{E}$
χ^2 independence test	×	A7, A8[b]	categorical, paired	T3	$\sum \frac{(O-E)^2}{E}$
McNemar's test	×	A7, large N_{12}, N_{21} counts	categorical, paired	T4	$\frac{(N_{12}-N_{21})^2}{N_{21}-N_{21}}$
Fisher's exact test	×	small sample, A5	categorical	T3	$\frac{N_{1+}!N_{2+}!N_{+1}!N_{+2}!}{N_{11}!N_{12}!N_{21}!N_{22}!n!}$
Wilcoxon signed-rank test	×	A7, symmetric	numerical, paired	T4	$\sum_{i=1}^{n_r} \text{sign}(Z_i) \cdot \text{Rank}(\lvert Z_i \rvert)$
Mann-Whitney U test	×	A5, A7	ordinal	T4	$\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{I}\{X_i > Y_j\}$
Friedman test	×	A7, A8	ordinal	T2[c]	$\frac{12}{ns(n+1)}\sum_{j=1}^{s}\left[R_{+j} - \frac{n(s+1)}{2}\right]^2$
Sign test	×	A7	paired	T4	TBD[d]
Bootstrap test	×[e]	A7, A8, A11	TBD	TBD	TBD
Permutation test	×	A6	TBD	TBD	TBD

[a] Also assumes expected counts > 5

[b] Also assumes expected counts > 5

[c] This is an approximate hypothesis. The official hypothesis is that there is no treatment effect in all systems.

[d] To be determined in each case.

[e] Parametric bootstrap exists.

Chapter 5

TEST COMPARISON AND CASE STUDY

This section will discuss the proper procedure of conducting a statistical significance test in NLP system performance comparison. We will also illustrate the procedures in multiple case studies which differ in the data type under consideration. The first case demonstrates a simple two-sample problem with simulated normal data. The second case includes a matched pair design of normal data. The third study examines a paired two-sample problem of dependent observations, using the data from 2017 WMT outputs. The fourth case pertains to a contingency table where we employ the McNemar's test.

5.1 Testing and reporting

The NLP system performance comparison starts with identifying what NLP task is under consideration and what kind of evaluation metric is used to measure the goodness of the system performance. It is necessary to understand whether the evaluation metric used is computed with respect to a single test instance or a set of test instances. For example, a sentence-level $BLEU$ score can be computed from a single system output, whereas an accuracy can be computed using multiple test instances–a set of tokens processed by a POS tagger which will produce one accuracy measure. In the latter case, a test instance is a single token; an evaluation unit is a set of tokens used to compute one accuracy measure. To simplify, an evaluation unit corresponds to a single value of evaluation metric.

Then, a hypothesis must be formed. Declare a hypothesis in regards to what is of interest: the mean (average performance), the median (a more robust measure of average performance) and distributional difference, etc. Alternatively, one may be interested in knowing the relationship between two systems; in such case, a hypothesis can be phrased with respect to the

Table 5.1: General testing and reporting procedure.

- Identify the NLP task and choose an appropriate evaluation metric.
- Form a hypothesis and choose a test statistic.
- Gather the data and conduct exploratory data analysis.
- Choose an appropriate significance test based on assumption checking.
- Reject or fail to reject the null hypothesis based on the obtained p-value.
- Report the sample size, the significance level, the obtained p-value, the significance test used and effect size.
- If necessary, conduct post hoc power analysis.

presence or absence of statistical independence (existence of association). In formulating a hypothesis, one must determine and clarify the test statistic used, the directionality (one-sided or two-sided) of the hypothesis and the significance level, etc. Also, clearly state the hypothesis (task).

In possession of the data, one must take the structure and sampling procedure of the data into account. In comparing two systems, the two samples are assumed to be paired; each sample can either contain independent or dependent observations, based on the fashion of sampling. Conduct pre-exploratory analysis (no p-hacking) of the data with respect to the normality and empirical distribution (histograms).

Subsequently, determine the significance test for system performance comparison. If the pre-exploratory analysis concludes that the samples are not significantly different from being

normally distributed and assumptions of the t test are verified, one can safely choose to use the paired student t test, given that the mean difference is of interest and the samples are numerical. However, if non-normality is detected, a nonparametric test may be more appropriate. Above all, verify the assumptions of the test.

After conducting a significance test, a p-value is obtained. One must draw the conclusion with respect to the original hypothesis tested. If the hypothesis pertains to the mean difference, then the appropriate conclusion drawn, given the p-value is smaller than the designated significance level, is that the null hypothesis is rejected and there exists enough evidence to suggest that the two systems perform statistically differently (in a two-sided hypothesis). If the p-value is greater than the significance level, do not accept the null hypothesis but rather state that at this moment the data suggests that we cannot reject the null hypothesis and find no statistically significant difference between the average performance of the two systems.

Depending on the hypothesis tested, choose an index of effect size and calculate the effect size. Report both the p-value and the effect size, and give a qualitative interpretation of the result. Time or space permitting, a power analysis of the test can be performed to better report the test result.

5.2 Case 1: simple 2-sample numerical data

In this case study, we consider a 2-sample numerical problem where the two samples are unpaired and independent. Given two samples $\mathbf{X}$ and $\mathbf{Y}$. Suppose that we are interested in knowing how on average the two samples differ from each other. Thus, the mean is the proper choice of statistic to compare.

5.2.1 The data

Since the NLP system performance tasks usually require matched pairs as the data, we will use simulation to demonstrate this case study. We simulate 100 observations in each sample, independently from a normal distribution with mean 0 and standard deviation 1 and from

a normal distribution with mean 1 and standard deviation 3. Denote the two samples by $\mathbf{X}$ and $\mathbf{Y}$. We have $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim N(1,3)$ where $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$.

5.2.2 Hypothesis

Without seeing the test results, we give the following hypothesis testing problem: on average $\mathbf{X}$ is equal to $\mathbf{Y}$, which is a two-sided hypothesis (Task 1).

$$H_0 : \mu_X = \mu_Y \quad v.s. \quad H_1 : \mu_X \neq \mu_Y \tag{5.1}$$

5.2.3 Test comparison

The following tests are all appropriate for this hypothesis testing problem regarding the mean difference: student t test, the bootstrap test for independent samples and the permutation test. If we are interested in investigating median differences, then the Mann-Whitney U test is an appropriate choice, or when the underlying distribution is known to be symmetric, in which case the median is the mean.

Assumptions

These four tests have different assumptions on the data. The student t test assumes that the two samples are unpaired and are both from a normal distribution and that each sample contains *i.i.d.* observations.

The bootstrap test assumes representativity and independence of the samples, and each sample has *i.i.d.* observations.

The permutation test assumes that two samples are independent (unpaired) and that observations in each sample are exchangeable. Also, the permutation test has a null hypothesis that two samples are from the same distribution.

The Mann-Whitney U test assumes independent samples, each containing independent observations; further, this test assumes that the samples are ordinal or at least can be meaningfully ranked.

Table 5.2: Test comparison in case 1.

Test name	**Hypothesis**	**Assumptions violated**	**Applicability**
Unpaired t test	T1	None	✓
Paired t test	T1	A5	×
Wilcoxon	T4 (median)	A5	×
Sign test	T4	A5	×
Bootstrap	T1	None	✓
Permutation	T2	None	✓
Mann-Whitney	T4 (median)	None	✓

Note that the four aforementioned tests have different null hypotheses. The t test and the bootstrap test are testing for mean equality, while the permutation test is testing for identical distribution. The Mann-Whitney U test is testing for median being equal to 0. Under the simulation scheme in this case study, we know that the underlying distributions of the two samples are normal, of which the mean and median correspond. Thus, the t test, the bootstrap test and the Mann-Whitney U are testing for the same thing. For the permutation test, the mean difference is used as the test statistic. Recall that in testing for identical distribution, rejection of the null implies stochastic dominance, which further implies mean difference. Thus, all four tests are equivalent in this case study.

Parametricity

Regarding parametricity of tests, the student t test is a parametric test which assumes that both samples are normally distributed, while the bootstrap test and the permutation test

are both nonparametric tests.

5.2.4 Verifying assumptions

For the student t test, normality can be verified by conducting a normality test, such as the Shapiro-Wilk test, or by graphical inspection using the Q-Q plot. The *i.i.d.* assumption is met in the collection process of the data, where each sample must be collected in a random fashion.

The assumption of independence between the two samples is also met during the data collection process, where the two samples are not observed in a paired setting.

Representativity assumption can be examined by looking at the spread of the histogram; the exchangeability assumption is satisfied if the two samples contains *i.i.d.* observations, since *i.i.d.* implies exchangeability but not vice versa.

Ordinality of the samples can be seen from the property of the evaluation metric, where we prefer larger values of the evaluation metric.

5.2.5 System performance comparison

To avoid *p-hacking*, we will simply choose the student t test as our testing method, since we know a priori that the two simulated samples come from normal distributions. Thus, we will use the unpaired student t test with unequal variance adjustment in this case. Given a significance level $\alpha = 0.05$, the p-value reported is 0.001, far smaller than α, indicating a significant difference between the population means of the two samples. The effect size defined by Cohen's d is **0.480**, indicating a fairly medium effect. Note that the Hedges' g (**0.478**) in this case is very close to the Cohen's d, since the sample size is quite large ($n = 100$). Here we will report the test results of all applicable tests mentioned before. In Table 5.3, the column of applicability refers to whether or not the assumptions held by the test are satisfied. It can be seen that all tests agree to reject the null hypothesis, which is known to be false.

Table 5.3: Testing result of case 1 (we expect a rejection).

Test name	P**-value**	**Applicability**	**Rejection**
t test	< 0.001	✓	✓
bootstrap	0.001	✓	✓
permutation	0.002	✓	✓
Mann-Whitney	< 0.001	✓	✓

5.2.6 Post exploratory data analysis

The histograms of sample **X** and **Y** are given in Figures 5.1 and 5.2. Note that both samples' histograms are symmetric and apparently normal, with no obvious outliers.

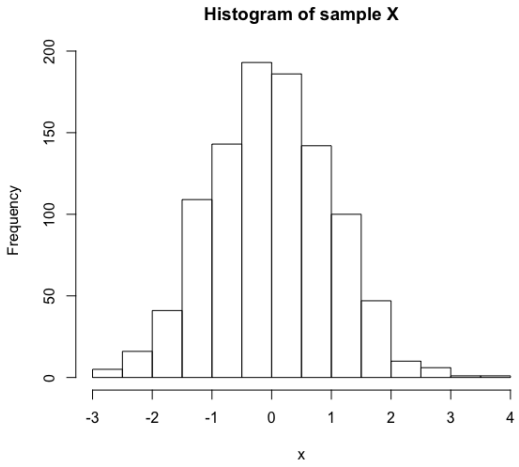

Figure 5.1: Histogram of sample **X**.

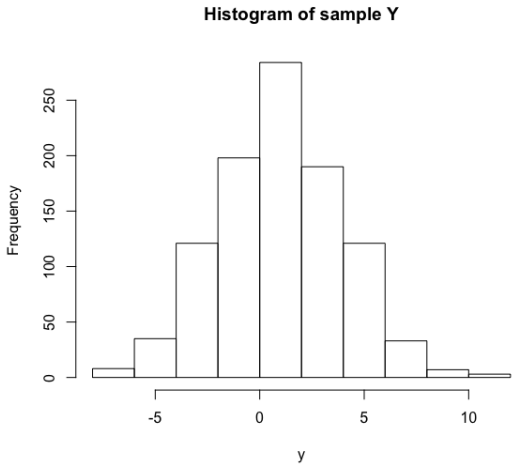

Figure 5.2: Histogram of sample **Y**

The sample mean difference is -1.006. Summary statistics of both samples are given in Table 5.4.

Table 5.4: Summary statistics for samples **X** and **Y**.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
X	-2.845	-0.681	-0.010	-0.003	0.682	3.589
Y	-7.807	-0.942	0.956	1.002	2.969	11.290

5.2.7 Power analysis

Here, we will compare all mentioned applicable tests in this case with respect to their statistical power via simulation. Since we know the exact distribution of the two samples, we can use Monte Carlo simulation for our purpose. We will invoke Algorithm 5 to simulate the power of the five tests compared in this case study, with different sample sizes from 2 to 200. Since $\mu_x \neq \mu_Y$, we are operating under the alternative and know that the alternative is true. The power curve is shown in Figure 5.3. Varying the sample size of the simulated data, we can see that generally the power grows to 1 as the sample size grows, for all tests. For all the tests, a sample of roughly under 100 observations will suffice to achieve a power level of 0.8. Note that the Mann-Whitney U test produces slightly smaller power than other tests from 50 to 130 in sample size. However, all tests approximately perform equally in this simulation, in spite of the fact that the assumptions of some tests are not satisfied.

5.3 Case 2: paired 2-sample numerical data

In this case study, we consider a 2-sample numerical problem where the two samples are paired and thus dependent; however, the two samples contain *i.i.d.* observations. Given two samples **X** and **Y** which are assumed to be paired. Suppose that we are interested in knowing how on average the two samples differ from each other. Thus, again the mean is the proper choice of statistic to compare. Task 1 is tested in this case.

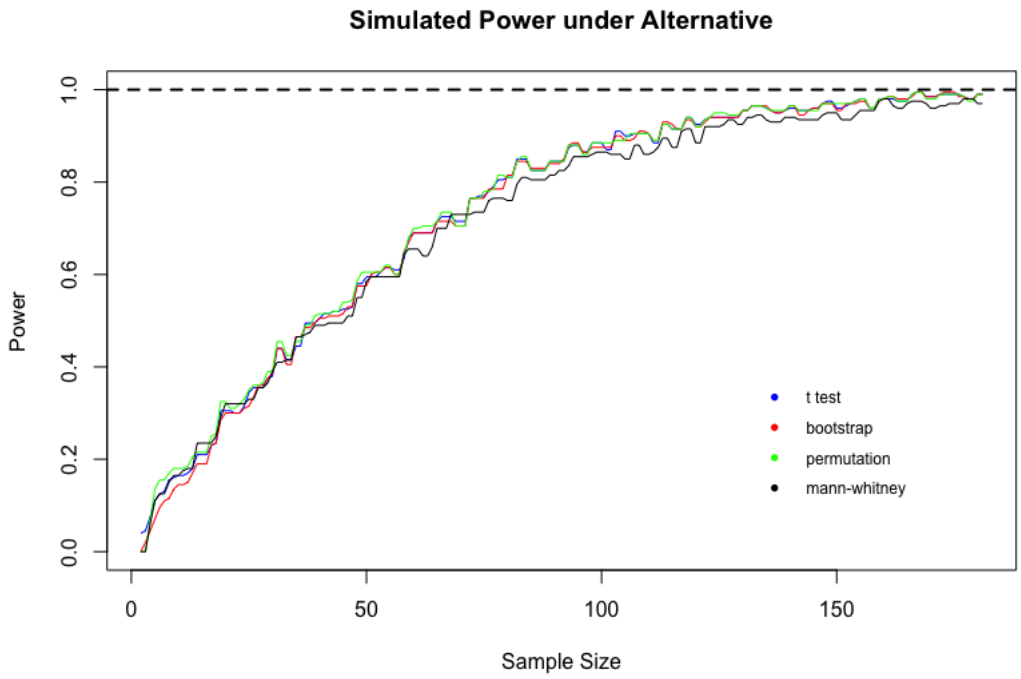

Figure 5.3: Simulated power curves for case 1.

5.3.1 The data

We again will use simulated data in this case study. To simulate paired data, we first simulate 100 *i.i.d.* observations according to an exponential distribution with intensity 0.5, which has a mean 2. For each observation, we add perturbation which is distributed according to a Cauchy distribution with location 0.5 and scale 1. Denote the two samples by $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{X} \sim \text{Exp}(0.5), \ \mathbf{Y} \sim \mathbf{X} + \text{Cauchy}(0.5, 1) \tag{5.2}$$

where the distributions of the perturbation and $\mathbf{X}$ are independent. The reason for using a Cauchy distribution is due to the fact that this distribution has a heavy tail. The choice of the exponential distribution is to make the sample non-normal. This construction will have an impact on the performance of parametric tests which rely on normality assumption and

tests which are sensitive to outliers.

5.3.2 Hypothesis

Without seeing the test results, we give the following hypothesis testing problem: on average **X** is equal to **Y**, which is a two-sided hypothesis (T1).

$$H_0 : \mu_X = \mu_Y \quad v.s. \quad H_1 : \mu_X \neq \mu_Y \tag{5.3}$$

5.3.3 Test comparison

The following tests are considered for this hypothesis testing problem regarding the mean difference: paired student t test, the bootstrap test and the sign test. The Wilcoxon signed-rank test is appropriate when the hypothesis is phrased with respect to the median or when the underlying distribution is symmetric, which in this case is true since the perturbation is Cauchy.

Table 5.5: Test comparison in case 2.

Test name	**Hypothesis**	**Assumptions violated**	**Applicability**
Unpaired t test	T1	paired data	×
Paired t test	T1	None	✓
Wilcoxon	T4 (median)	None	✓
Sign test	T4	None	✓
Bootstrap (paired)	T1	None	✓
Permutation (unpaired)	T2	paired data	×
Mann-Whitney	T4 (median)	paired data	×

Assumptions

In the paired student t test, the assumption of paired data is satisfied since we simulate the data in this fashion. The assumption of *i.i.d.* observations is also satisfied due to random sampling. Normality assumption is violated since we sample from an exponential and a Cauchy distribution. Therefore, the paired student t test is not applicable in this case to detect mean difference.

The bootstrap test can be employed with a null hypothesis that the true mean difference is 0. The resampling method will respect the matched pair nature of the data, which suggests that we resample each pair in each iteration. The representativity assumption will be assumed and the *i.i.d.* assumption is satisfied due to random sampling.

The sign test is also applicable in this case, since this is a matched pair design. Also, the *i.i.d.* assumption is satisfied due to random sampling.

In the Wilcoxon signed-rank test, the assumptions of paired data and *i.i.d.* observations are satisfied. Note that the two original distributions of the samples are not symmetric, but the perturbation is symmetric around 0.

Parametricity

The paired student t test is a parametric test, which assumes normal distribution on the data. The remaining proposed tests are all nonparametric. Due to the fact that we know the distribution of the samples is not normal, the student t test's assumption is violated. All other candidate tests are nonparametric tests.

5.3.4 Verifying assumptions

In this case study, we will use the sign test (two-sided). Again, the sign test assumes *i.i.d.* observations and paired data, which are both satisfied in this case study.

5.3.5 System performance comparison

Let the significance level be 0.05 and iteration time be 2000. We use the mean difference as the test statistic and reject for large absolute value of the permuted mean difference. The two-sided sign test outputs a p-value of 0.013, implying a rejection of the null hypothesis. The Cohen's d and Hedges' g are given by **0.291** and **0.288**, which are very close and suggest a fairly small effect. Given the way of data simulation, we know that the null hypothesis is false, since on average the true shift of mean is positive (0.5). Here we report all test results of aforementioned tests. Note that only the sign test and the Wilcoxon test are able to definitely reject the null, whereas the bootstrap test based on t ratios cannot reject the null, and the t test has a p-value that is very close to 0.05.

Table 5.6: Testing result of case 2 (we expect a rejection).

Test name	***P*-value**	**Applicability**	**Rejection**
sign test	0.002	✓	✓
paired t test	0.04	×	✓
bootstrap	0.194	✓	×
Wilcoxon	< 0.001	✓	✓

5.3.6 Post exploratory data analysis

The histogram (Figure 5.4) of the difference $\mathbf{Z}$ is left skewed with an obvious outlier around -120. The empirical distribution has a mean slightly smaller than 0 with a majority of data clustering around 0.

5.3.7 Power comparison

Here, we will invoke the power simulation algorithm to simulate statistical power for all proposed tests in this case study. Figure 5.5 displays four simulated power curves for the

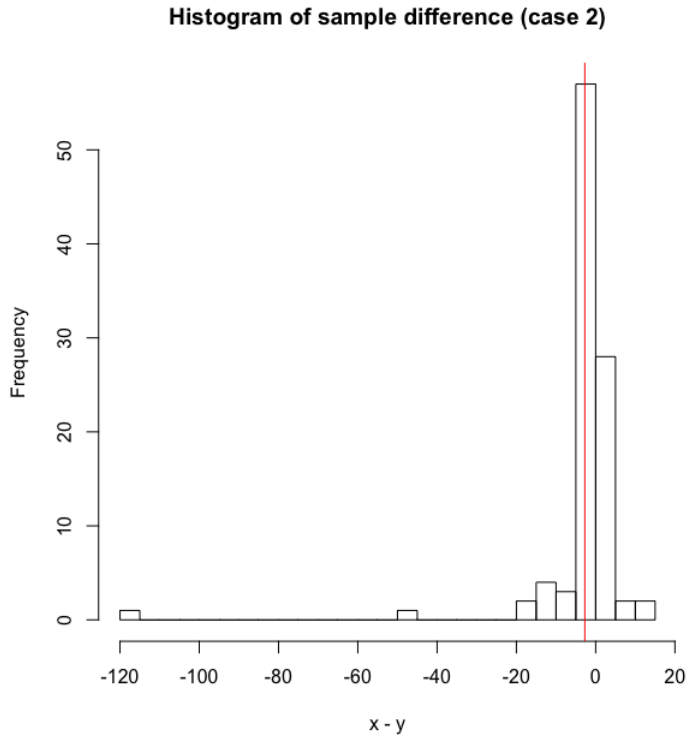

Figure 5.4: Histogram of difference **Z**.

paired t test, the bootstrap test, the sign test and the Wilcoxon signed-rank test.

This plot illustrates how the power of a test is associated with different sample sizes. It can be seen that for the bootstrap test based on the median and the sign test based on the median, to achieve a power level of 0.8, we only need approximately a sample of size 100. For the Wilcoxon test, a sample of 150 observations will be required.

Based on our construction of the simulation, where we impose that the true mean difference is not 0, only the Wilcoxon signed-rank test is able to detect such difference, while other aforementioned tests fail due to constantly low power as the sample size grows. In contrast, it should be noted that the paired t test has low simulated power, because of the violation of normality and the presence of outliers due to the heavy-tailed distribution of the perturbation, where as the Wilcoxon signed-rank test is not affected by outliers since it is a

Table 5.7: Summary statistics for samples **X**, **Y** and **Z**.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
X	0.030	0.431	1.079	1.789	0.682	2.624
Y	-12.176	0.509	1.957	4.524	4.210	117.691
Z	-117.654	-1.977	-0.578	-2.735	0.225	14.140

rank-based test.

In the bootstrap test, we use the t ratios to calculate a confidence interval to make inference on the mean difference. The choice of the t ratios is also sensitive to outliers, which affects the performance of the test. In the sign test, we use the mean difference as the test statistic as well. Since the mean is extremely sensitive to the presence of outliers, it makes sense that the bootstrap and sign test fail to detect the true location shift in this case. Note that the Wilcoxon signed-rank test has a different null hypothesis that the distribution of the true difference is symmetric around 0 (or the median is 0), which is obviously false since the perturbation has median (and mode) being 0.5; further, the mean of the Cauchy distribution does not even exist. Hence, the Wilcoxon signed-rank test outperforms the paired t test, the bootstrap test and the sign test in this case. Furthermore, this case illustrates the core difference between the student t test and the Wilcoxon signed-rank test, the latter of which is usually said to be the nonparametric alternative to the former. The paired student t test makes inference on the population mean, which determines that its null hypothesis is phrased with respect to the mean difference. The Wilcoxon signed-rank test makes inference on the population median, which is equivalent to the paired student t test only when the underlying distribution is known to be symmetric. Otherwise, the student t test's assumption will be violated, since normality presupposes symmetry.

In addition, we reapplied the bootstrap test and the sign test using the median of the difference as the test statistic, where the null hypothesis is to test for symmetry around 0. In this case study, the distribution of the sample difference is Cauchy, which means that

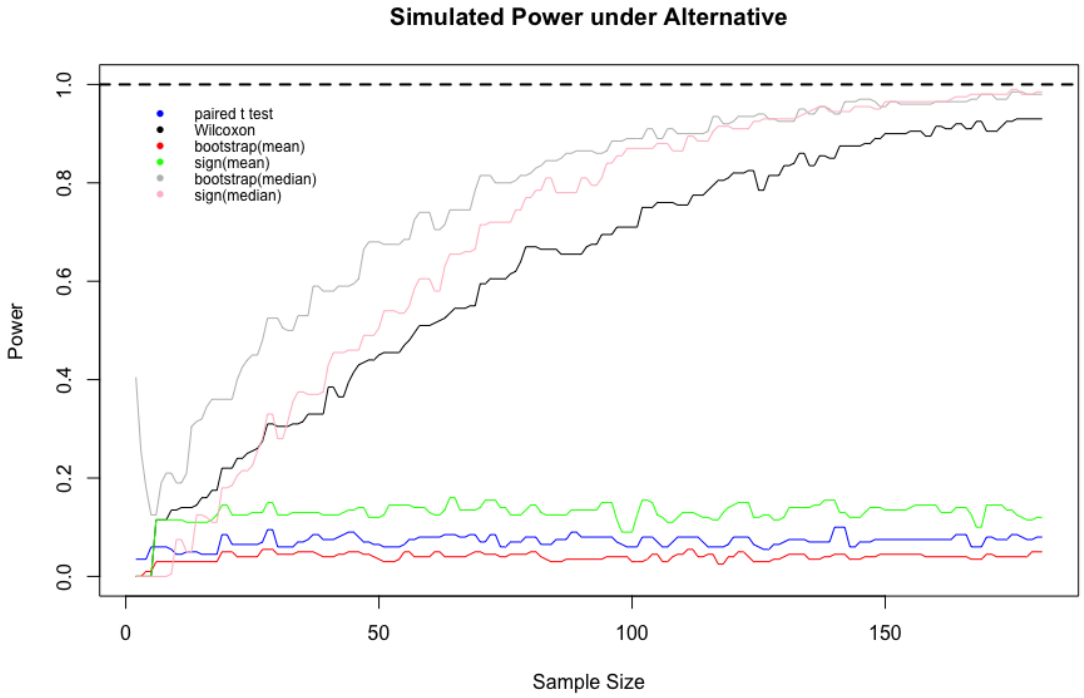

Figure 5.5: Simulated power curves for case 2.

the median and the mode is the location parameter, which is 0.5. In the bootstrap test, we construct a confidence interval based on the observed median of the difference. Let $\hat{\theta}$ be the observed median of difference. The confidence interval is given by

$$[2 \times \hat{\theta} - \theta^b_{\alpha/2}, 2 \times \hat{\theta} - \theta^b_{1-\alpha/2}] \tag{5.4}$$

where θ^b is the bootstrapped empirical distribution of $\hat{\theta}$. This bootstrap confidence interval is called the basic bootstrap confidence interval (Davison and Hinkley, 1997).

The sign test using the median difference as the test statistic is analogous to the one using the mean difference, and the p-value is obtained in the same fashion.

It is obvious that by using the median difference as the test statistic, the two tests achieved even higher power than the Wilcoxon signed-rank test. This case illustrates how the student t test is sensitive to outliers and how the choice of a test statistic can influence

Table 5.8: Additional testing result of case 2 (we expect a rejection).

Test name	P**-value**	**Applicability**	**Rejection**
bootstrap (median)	0.001	✓	✓
sign test (median)	< 0.001	✓	✓

the inference. It is important to conduct exploratory data analysis on the samples before choosing a test.

5.4 Case 3: dependent numerical samples

In this case study, we consider paired and dependent two-sample numerical data obtained from the 2017 Machine Translation shared task (Bojar et al., 2017). We will choose any two system outputs and compare their performance on the test set. In this particular study, we will apply the block bootstrap method.

5.4.1 The data

The published test set for Czech-English news translation, in the Standard Generalized Markup (SGM) format, contains 3005 sentences, which constitute 113 news articles. The candidate machine translation system takes each sentence as its input and produces a hypothesis translation, which will be evaluated with respect to the sentence-level $BLEU$ score against a reference sentence. We will use two systems named *online-A.0.cs-en* and *online-B.0.cs-en*, denoted by A and B respectively in this case study.

Here, we identify elements of the comparison tasks defined in Chapter 2. In this comparison task, the test instance is a single sentence in the test set. The test set is a set of 3005 sentences, and there are 113 news articles. The evaluation unit is a test instance, from which a $BLEU$ score can be calculated with respect to its corresponding gold standard. The evaluation metric is the sentence-level $BLEU$ score. Note that the choice of the sentence-level $BLEU$ score can greatly impair the evaluation of translation quality. We use the sentence-

level $BLEU$ here only for illustrative purpose. For a more precise and meaningful translation quality measure, Collins, Koehn, and Kučerová (2005) suggested an approximation of per sentence score to measure translation quality.

In particular, we can notice that there exist dependency structures in the test set: sentences are from articles, which indicates that the test set does not contain sentences due to simple random sampling. Instead of assuming generally that each sentence is independent of the others, we assume that each article is independent of other articles, and sentences within each article are dependent, since given the presence of one sentence in a particular article, other sentences in the same article will surely appear as well, which violates the definition of independent samples. Therefore, based on the structure of the test set, we consider using the block bootstrap as the calibration of this hypothesis testing problem.

5.4.2 Hypothesis

Without seeing the test results, we give the following hypothesis testing problem: system A outperforms or performs as well as system B in the given test set. As constructed before in this paper, the corresponding hypotheses with respect to the average of values of the BLEU scores are as follows.

$$H_0 : \mu_A \geq \mu_B \quad v.s. \quad H_1 : \mu_A < \mu_B \tag{5.5}$$

where μ represents the population mean of the distribution of the $BLEU$ scores for each system. We decide to use a one-sided hypothesis because we are interested in knowing which one performs better instead of different performance.

5.4.3 Test comparison

Note that the samples are paired and both contain dependent observations. Among the various tests we have introduced, the block bootstrap appears appropriate in this case. However, ignoring potential dependencies within the test set, for comparing population central tendency (mean and median) differences, the paired student t test, the Wilcoxon signed-rank

test and the sign test calibrated by permutation can be used.

Table 5.9: Test comparison in case 3.

Test name	**Hypothesis**	**Assumptions violated**	**Applicability**
Unpaired t test	T1	paired data, T7	×
Paired t test	T1	T7	×
Wilcoxon	T4 (median)	T7	×
Sign test	T4	T7	×
Bootstrap (paired)	T1	T7	×
Permutation (unpaired)	T2	T7	×
Mann-Whitney	T4 (median)	T7	×
Block bootstrap	T1	None	✓

Note that in this case study, the evaluation metric is the *BLEU* score, which is a real number from 0 to 1. The student t test assumes normality for the test to work properly; however, normal distributions range over the entire real number. One could consider a student t test based on truncated normal distribution instead to avoid this difference in domains upon which the BLEU score and the underlying probability distribution are defined.

5.4.4 Verifying assumptions

Since we are using the block bootstrap, we assume independence among blocks. Such independence assumption cannot be verified and can only be taken for granted; however, assuming independence among articles/blocks makes more sense than assuming independence among sentences directly. Also, we must assume that the observed BLEU scores are

representative of the population.

5.4.5 System performance comparison

Let the significance level $\alpha = 0.05$. We invoke Algorithm B.4 to implement the block bootstrap in this case, and we choose the studentized t ratio as our test statistic to make inference on the population mean. The t ratios are used to calculate the one-sided 95% confidence interval, which is given by

$$\big(-\infty, -0.010\big) \tag{5.6}$$

Note that 0 lies outside the confidence interval, which means that we can reject the null hypothesis that the two samples have the same mean. However, let's calculate the effect size using the Cohen's d as the index, which is 0.099, implying a trivial effect. Even though the test suggests a statistically significant result and rejects the null, the effect size is so small that it can be neglected.

Here, we also report the test results of all aforementioned tests. Note that all tests reject the null hypothesis based on very small p-values, but all tests except the block bootstrap test do not have their assumption satisfied in this case. The obtained significance may be merely due to large sample size. In

Table 5.10: Testing result of case 3 (the official report of the evaluation results gives a rejection).

Test name	***P*-value**	**Applicability**	**Rejection**
bootstrap (mean)	< 0.001	✓	✓
paired t test	< 0.001	×	✓
Wilcoxon	< 0.001	×	✓
sign test (mean)	< 0.001	×	✓

5.4.6 Post exploratory data analysis

After the test, we conduct a post exploratory data analysis on the dataset. The histograms for BLEU scores of systems A and B are given in Figure 5.6.

Note that in both histograms, the data range from 0 to 1, which is the range of the BLEU score. Both samples are right skewed, centering around 0.2. The histograms suggest that both samples may not be normally distributed. The sample size is 3005 for both samples, and there are 113 articles. Also, in the histograms of both samples, the overlapping portion (in purple) is quite large, consistent with the trivial effect size.

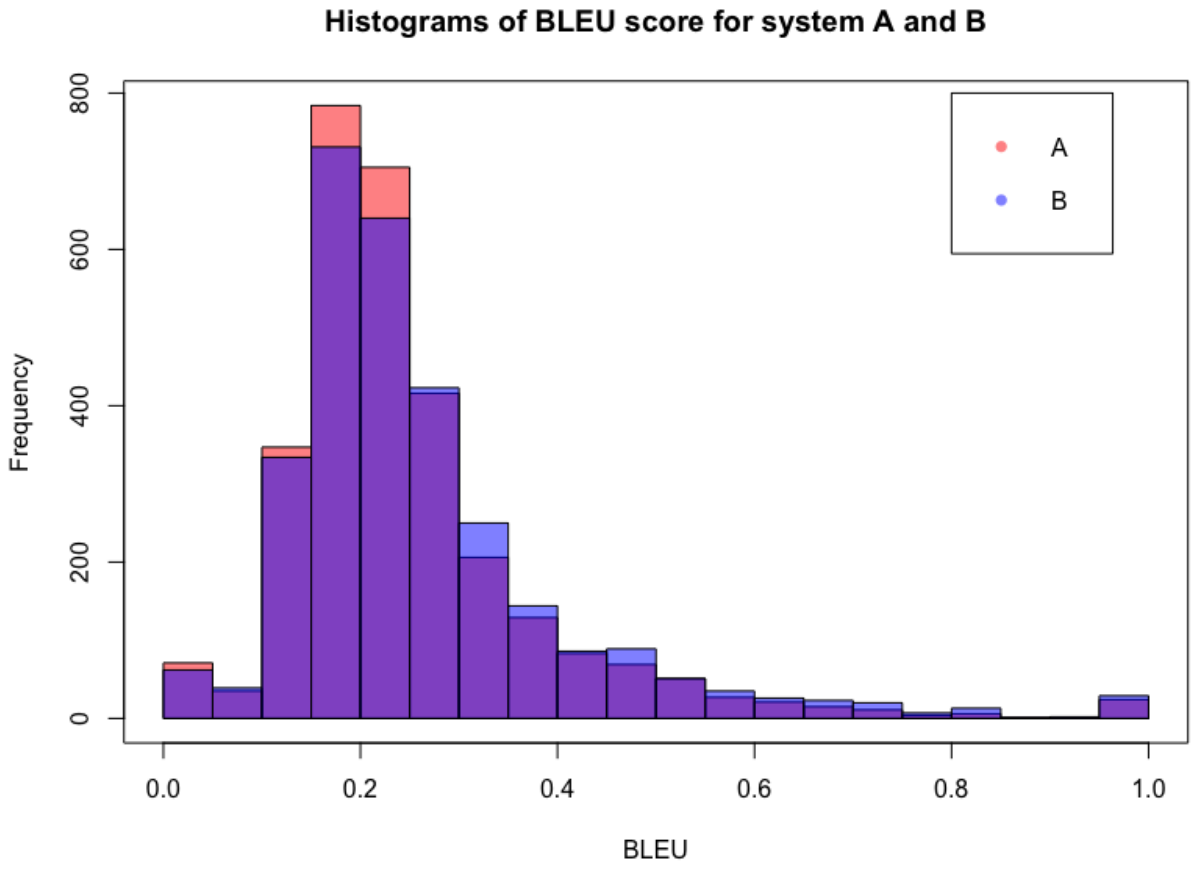

Figure 5.6: Histogram of $BLEU$ scores of system A and B.

The sample mean difference is -0.014 and the Cohen's d is 0.099, which implies a negligible effect. Summary statistics of both samples are given in Table 5.11.

Table 5.11: Summary statistics for the given samples.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
A	0.000	0.172	0.218	0.249	0.284	1.000
B	0.000	0.174	0.227	0.263	0.305	1.000

Now, we apply other statistical tests to this dataset to examine the same set of hypotheses simply for comparison. First, let's use the paired Welch's t test. We must note that the $i.i.d.$ assumption is unfortunately violated due to the data structure existing in the test set. The normality assumption can be ignored since we have a large sample size. The paired t test produces the same result ($p \ll 0.05$), suggesting that we have enough evidence to reject the null hypothesis and conclude that there is statistically significant difference between the two samples, and thus system A underperforms system B on average given the test set. However, note again that the effect size Cohen's d is equal to 0.099, suggesting a very small effect, and thus even if the difference is statistically significant, the effect is negligible.

Now, we turn to the Wilcoxon signed-rank test. Again, the independence assumption among observations is violated. This test also produces the same result as the t test, suggesting a rejection as well.

The 2017 WMT evaluation procedure (Bojar et al., 2017) is done similarly, but with a different evaluation metric, where the translation quality is based on the so-called Direct Assessment (DA) standardized mean adequacy scores whereas in this case study we use the sentence-level $BLEU$ score as the evaluation metric. The statistical test used is the Wilcoxon signed-rank test. The test is based on the significance level of 0.05 and the test result suggests that system A underperforms system B significantly. Although the official evaluation result is consistent with our result using the same test, note that the assumption is not satisfied in the first place; thus using the Wilcoxon signed-rank test is invalid. Also

note that this case study and the official report do not use the same evaluation metric, for which the conclusions drawn could either differ or agree.

Also, the official report does not mention p-value adjustment while numerous pairwise comparisons between systems have been done. Most importantly, even though all tests report statistical significance, the effect size calculated by the Cohen's d suggests a negligible effect. The significance of the reported p-values can be seen from the large sample size ($n = 3005$). As aforementioned, with large sample sizes, tests tend to be statistically significant, despite trivial effects. Therefore, merely relying on the reported significance or p-values when the sample size of the test set is large is not sufficient for drawing conclusions on the superiority of the system performance. Along with the obtained p-values, predetermined α level and the sample size of the test set, the effect size should also be reported and taken into account. In conclusion, the statistical significance test has found significant difference between the performance of the two systems, but the effect is negligible. This kind of apparently paradoxical conclusion submits to an over-powerful study, where the sample size is large enough so that any trivial discrepancy tends to be magnified by the significance test.

5.4.7 Power comparison

Here, we will invoke Algorithm 6 to simulate the statistical power for the tests mentioned above. We will repeatedly draw observations from the two samples with replacement, which form the bootstrapped samples. Then, we add the Cohen's d to the bootstrapped samples of system B in order to operate under the alternative. By adjusting the $BLEU$ scores of system B by the amount of the estimated Cohen's d (≈ 0.099) and applying all candidate tests using different sample sizes, the power curves are given in Figure 5.7. The power curves of this case demonstrate a quite strange phenomenon: all tests exhibit a power level of 0.8 after a bootstrapped sample of size 10, after which all tests have approximately the same power, where the block bootstrap tends to have lower power as the sample size grows. Note that there is a surge of power before the sample size reaches 10.

This surge of power may be due to many reasons. First, the sentence-level $BLEU$ score

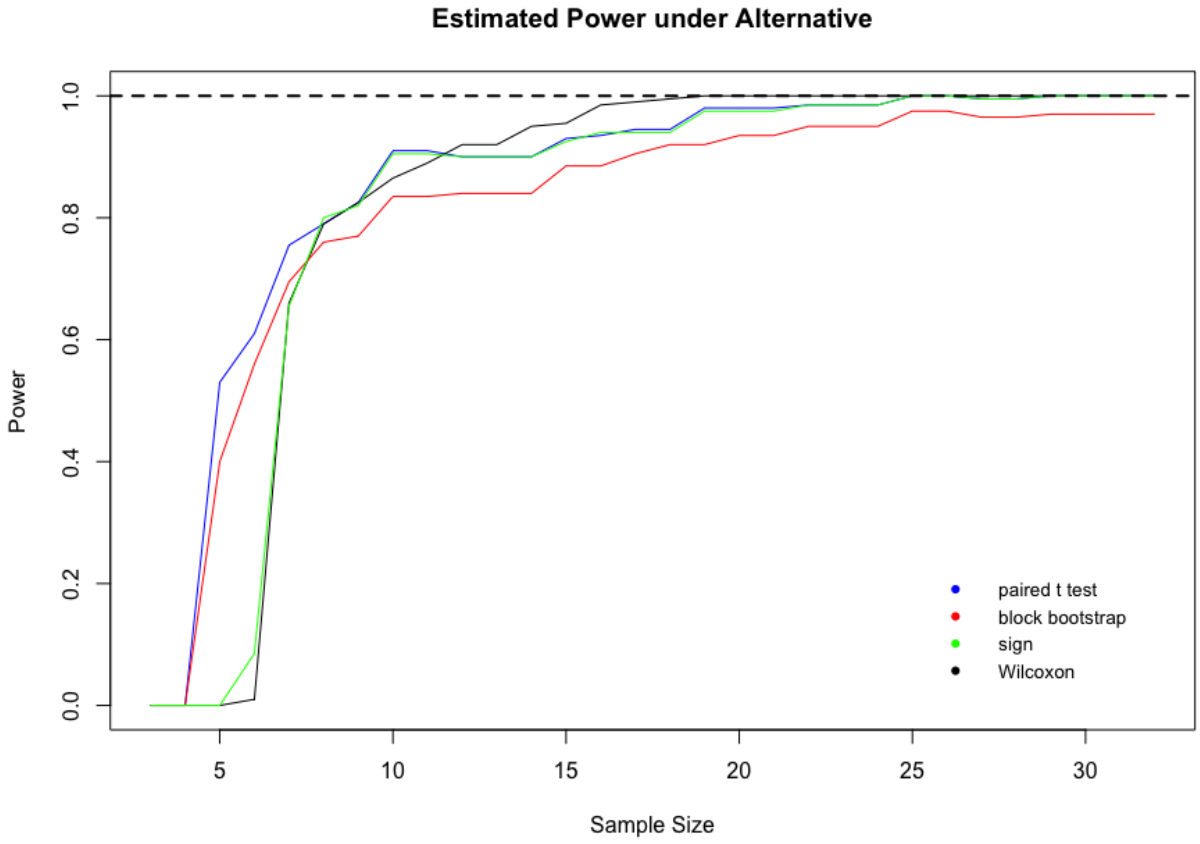

Figure 5.7: Estimated power curves of case 3.

is low in efficiency when the sentence length is short, resulting in hardships in comparing the overlapping n-grams of the two sentences, which may further cause the existence of ties of $BLEU$ scores even though they are in several decimals. The existence of ties tremendously prohibits the use of rank-based tests such as the Wilcoxon signed-rank test and the Mann-Whitney U test. Sometimes the existence of an excessive amount of ties also renders the t test useless since the sample variance may be 0 when the sample size is small. The second reason might reside in the fact that the two systems indeed are different, and the tests can detect such difference when the sample size is small. However, it still appears bizarre that the power soars when the sample size reaches 10. Nevertheless, one conclusion that can be

drawn from the power analysis is that the use of Wilcoxon signed-rank test in the original paper based on the size of the original test set appears overpowered, meaning that if the desired power level is 0.8, a sample of size roughly over 10 will suffice, whereas the actual sample size surpasses the minimally required one sunstantially.

5.5 Case 4: 2-sample categorical data (contingency table)

In this case study, we will consider a 2-sample categorical data problem. We will use the test data of two systems: *dej*00 (Déjean, 2000) and *hal*00 (Halteren, 2000), from CoNLL 2000 chunking shared task.

5.5.1 The data

The test set used in this shared task contains 49389 tokens. The original test set consists of sentences, and the task is to tag each token with either B, meaning the beginning of a chunk, I, meaning being inside of a chunk or O, meaning not belonging to any chunks. It can be seen that the probability of having an I before B is quite large, and the probability of having a B after O is always 1. Thus, there might exist dependencies among the samples. However, the test for paired categorical data with dependent observations is not discussed here. For simplicity, we will assume that the observations are independent in this case study.

Here, we attempt to compare the test results of the two chunking systems, for which this comparison task is a matched pair design. First, we will compare the test system output of each system with the gold standard and obtain a vector of 1 and 0, where 1 indicates correct prediction and 0 indicates wrong prediction. Based on the two vectors of 1 and 0 for the two systems, we can form a 2×2 contingency table, where the rows are given by system *dej*00 and columns are given by system *hal*00. The contingency table is given in Table 5.12.

5.5.2 Hypothesis

In this case study, we wish to investigate whether the two systems have the same proportion of correctly predicted chunks. To properly form the hypothesis, we now turn to Table 4.4.

Table 5.12: Contingency table for case 4

	0	1
0	1463	994
1	476	46456

The number of correct prediction for system $dej00$ is given by $N_{11}+N_{12}$ and that for system $hal00$ is given by $N_{11}+N_{21}$. Note the common term and thus we only need to compare N_{12} and N_{21}. If we consider the proportion of errors, it is the same comparison done on N_{12} and N_{21}.

$$H_0 : p_{12} = p_{21} \quad v.s. \quad H_1 : p_{12} \neq p_{21} \tag{5.7}$$

5.5.3 Test comparison

Due to the matched pair nature of this case, the McNemar's test is appropriate to test for marginal homogeneity in the two systems.

Table 5.13: Test comparison in case 4.

Test name	**Hypothesis**	**Assumptions violated**	**Applicability**
χ^2 GOF	T2	paired data	×
McNemar test	T4	None	✓

5.5.4 Verifying assumptions

The assumptions of McNemar's test are that the two samples are dependent and nominal (or categorical) and that observations in both groups only submit to one category. Here, the groups are the two systems and the categories are the correctly tagged chunk and incorrectly

tagged one, coded by 1 and 0. Here, in a system, a test instance can only be tagged correctly or incorrectly, not both, which satisfies the assumption.

5.5.5 System performance comparison

By conducing McNemar's test, the obtained p-value is far less than 0.05 ($\ll 0.05$), suggesting a rejection of the null hypothesis that the two systems have the same proportions of discordance, which further implies that the proportions of correct predictions for the two systems are statistically significantly different. This is a two-sided test, and thus this is the only conclusion we could draw. However, by inspecting the sample proportions, we can see that the correct prediction of $dej00$ is smaller than that of $hal00$.

Table 5.14: Testing result of case 4 (the official report of the evaluation results does not use significance testing).

Test name	***P*-value**	**Applicability**	**Rejection**
McNemar test	< 0.05	✓	✓

Table 5.15: Qualitative interpretation of odds ratio and Cohen's g

Index	**Effect size**	**Interpretation**
Cohen's g	$(0.05, 0.15]$	small
	$(0.15, 0.25]$	medium
	$(0.25, \infty)$	large
Odds ratio	$(1.22, 1.86]$	small
	$(1.86, 3.00]$	medium
	$(3.00, \infty)$	large

This testing is for categorical data in a contingency table. Appropriate choices of effect

size index are the odds ratio and Cohen's g. R has a package of calculating these indices [1]. The sample effect size using the odds ratio is 2.088 and the Cohen's g is 0.176, suggesting a medium effect. Again, the qualitative interpretation given in Table 5.15 according to Cohen (1988) is quite arbitrary.

[1] See `https://rcompanion.org/handbook/H_05.html`.

Chapter 6
DISCUSSION

In this chapter, we will first discuss a different perspective on the NLP system performance evaluation tasks partially implied in other papers, where the authors perform the hypothesis testing on the level of the system output. Secondly, we will talk about limitations of previous work with respect to the lack of statistical formulation, the proper use of the test and the necessary reporting of the test result. Thirdly, we will highlight important findings in the case and simulation studies. Finally, we will briefly touch on limitations of this thesis with some open issues related to the case studies.

6.1 Comparison with previous studies

6.1.1 Test instance and evaluation metric

In this thesis, we assume the values of the chosen evaluation metric convey all the information we need in order to carry out the system performance comparison task, by which we only need to operate on the level of evaluation metric. Dror et al. (2018) in their description of the statistical significance testing suggested that an evaluation metric is a statistic of the system output, for which the test statistic defined in their paper is the difference of the evaluation metric between two systems.

Such consideration may sound sensible in that the system outputs are regarded as the samples and the null hypothesis is that the two systems produce statistically identical outputs, based on the chosen evaluation metric. However, they did not define how an evaluation metric is computed from the system outputs, which tends to cause misunderstanding. For example, a machine translation system evaluated on a test set consisting of multiple sentences will generate the same number of sentences. For each proposed translation, there is a corre-

sponding $BLEU$ score. Thus, there are the same number of $BLEU$ scores as the proposed translation and the test set. In another example of POS tagger comparison, a test set of many sentences or articles is given, and the competing systems both generate predicted tag for each token. Using the accuracy as the evaluation metric, one accuracy will be computed from the entire system outputs, with respect to the total number of correct predictions. It is also possible that the test set encompasses multiple subsets of tokens/sentences, where each subset generates a value of accuracy. In this scenario, then, it is similar to the machine translation system performance comparison.

In the two examples above, the computation of the chosen evaluation metric is defined differently. The sentence-level $BLEU$ score is computed from a single sentence, which is a single system output, whereas the accuracy is computed from the entire system outputs. In the work of Dror et al. (2018), the discrepancy between testing for a single number of evaluation metric (accuracy) and for a sample of multiple observations ($BLEU$) is not clarified.

Of course it makes sense for system performance comparison tasks to operate on the system output level, by assuming the null hypothesis being that the two systems are not different and generate the same outputs. In this case, the samples (two sample of outputs) are also paired, with respect to one test instance. The goal is to find a statistic which quantifies the difference of the two outputs. Note that for almost all the NLP systems, the outputs are not pure numbers but rather words, sentences or even summaries, which pose even harder obstacles in formulating the performance comparison task into a statistical framework. First of all, what does it mean for natural language data to be independent and identically distributed? Independence can be understood the same way of random sampling: each system output is independently generated by the system from other outputs. identically-distributedness, however, does not show a clear statistical connotation for natural language data. For two sentences, they can be of different lengths; they can be on different topics; the syntactic structures could also differ. Secondly, in this construction, how to incorporate the comparison with the gold standard? Merely comparing two systems' outputs with each

other provides relative results in that the only sensible conclusion to be drawn is that the two systems' outputs display no statistically significant difference–that is the two systems are equivalent with respect to the given test set. What if the two systems are both 'bad', compared to the gold standard? It is not ideal to provide a relative comparison in NLP or other fields for a system performance comparison task.

Deeming the values of an evaluation metric as the samples, however, can be more meaningful. The values of the evaluation metric are pure numbers, and thus it is meaningful to discuss its probability distribution, if we are in a probabilistic model. Also, for each system, the values are in the same unit or measurement, which allows for an easier comparison and inference. If two sets of values of the same evaluation metric are tested to be identically distributed, then it can be inferred that based on this chosen metric, the two systems generate identical results, and thus the two systems perform equally. In addition, the gold standard is also used in the comparison. To calculate values of the evaluation metric, we compare the system outputs of the two systems with the gold standard, independently. Then, the comparison performed this way is relative to the gold standard instead, enabling more meaningful interpretation of the testing results.

6.1.2 Formulation of statistical hypothesis testing

Although statistical significance testing has become the reliable method for researchers to provide scientific judgments on the system performance, many papers do not provide or report the hypothesis (task) considered in their work. It is unclear whether the average performance is compared, the median performance is compared or the distributions in general are compared for example in Koehn (2004), Bisani and Ney (2004).

It is important to clarify the null hypothesis and the alternative hypothesis in a system performance comparison task. What kind of comparison is of interest is closely related to the choice of test statistic and to the phrasing of the hypothesis. If one is interested in comparing average performance of the two systems, then the mean is a proper test statistic to use. If one is worried that the systems might generate outputs which may produce outlier values or

that the distribution of the evaluation metric is skewed, then a more robust statistic can be considered, such as the median.

The choice of test statistic and formulation of the null hypothesis will also give rise to the conclusion drawn from the test. If the mean is the test statistic, then the corresponding conclusion must be drawn with respect to the mean or average performance. If one is interested in the distribution of evaluation metric, then the conclusion must be drawn with respect to the distribution itself. It is a usual case that many conclusions from statistical significance tests are phrased not with respect to the test statistic itself but in a more generalized manner in order to provide easier interpretation for a wider range of audience. For example, a comparison task done with respect to the mean on a single-domain test set concludes that the proposed system outperforms the baseline system. Such conclusion is of course easy to interpret but cannot be inferred from the test per se, without mentioning the test statistic, the domain of the test set and the significance level.

6.1.3 Test assumptions

As given in Chapter 5, we listed the test assumptions for all mentioned tests. Each test has its own assumptions, and it is important to verify whether the assumptions are met or not before using the test. Koehn (2004) explained the assumptions of the significance tests introduced in their paper and mentioned ways to verify the assumptions. Demšar (2006) also discussed in length the test assumptions and suggested discontinuing the use of the Wilcoxon signed-rank test accordingly. The discussion given by Demšar (2006) only applies to significance testing for classifiers. Yeh (2000) introduced many significance tests based on whether the independence assumption is made by the tests. Yeh (2000) carefully listed the assumptions of the student t test (Welch's) and of the chi-square test, without scrutiny into other significance tests in terms of test assumptions.

Verifying test assumptions is a less mentioned topic in the significance testing in NLP. In this thesis, we repeatedly stressed the important of verifying test assumptions before using a significance test. Although it has been seen in case study 2 that sometimes without satisfying

the test assumptions, a significance test may be able to reject the false null. However, the occurrence of rejection when the test assumptions are not met is post hoc–one can only make such conclusion based on post hoc examination and comparison of multiple significance tests, as done in case study 2, which is not a recommended standard practice in real-world testing scenario. The statistical significance tests are built upon their assumptions and one can only enjoy their theoretical advantage if their assumptions are met.

6.1.4 Effect size

The use of effect size in significance testing in NLP is rarely mentioned. Søgaard (2013), using meta-analysis, specifically examined the estimation of effect size across datasets in NLP in order to evaluate the robustness of NLP systems. This short paper did not discuss particular statistical tests, their presuppositions or their use in NLP. We in this thesis only considers the scenario of a single test set, which, in the language of Søgaard (2013), leads to a micro estimation of the effect size. We also introduced several indices to estimate effect size for different data types. Again, the effect size is a measure of existing phenomenon on the population level, independent from the sample size and the significance test used. Reporting effect size estimates of a system performance comparison experiment is essential to evaluating the existence of performance difference qualitatively.

6.2 The choice and use of significance tests

After formulating the appropriate null and alternative hypotheses and choosing a test statistic, a significance test should be selected to perform the comparison task. Having the data in hand, one should first conduct exploratory data analysis to have an initial understanding of the data, descriptively. Histograms, boxplots and other graphs may be used to accomplish such goal. Summary statistics will also help. In a histogram, one should pay attention to the spread and general shape of the samples. Do they look like normal distributions? Where are they centered at? What is the range of the two samples? Note that the theoretical range of some evaluation metric may be from 0 to 1, whereas a normal distribution ranges over the

entire real number.

To use a parametric test such as the student t test, one should check for normality first. If the normality test suggests no significant deviation from a normal distribution, then a parametric test which assumes normality may be more appropriate than a nonparametric test. In a large sample, normality assumption can be relaxed. If such distributional assumption is not met or one wishes not to conduct a test to check the distribution of the samples (a Pearson chi-square test of goodness of fit test can be used to do so, given a proposed distribution, but it lacks power), a nonparametric test can be considered.

The choice of a nonparametric test should first be based on what type of data are the given samples. Usually in NLP system performance comparison tasks, we face two major types: numerical and categorical (binary). Accuracy, F1 score, BLEU score, ROUGE score and other metrics are numerical data, while the binary classification of correct and wrong predictions is categorical. Note that the latter is an alternative way to F1 score and accuracy of evaluating classifiers such as the POS tagger and chunking systems.

Then, one should consider how the test set is collected. Does the test set contain independent observations/test instances due to random sampling with replacement? Are the test instances exchangeable? Are the test instances dependent? Note that the identically-distributedness cannot be verified directly but is rather assumed. Among nonparametric tests, one of the greatest assumption differences is in the assumption of the data: independent samples, paired samples, samples with independent observations, samples with exchangeable observations and so on. Different data structures (dependent, independent or exchangeable) give rise to the use of different nonparametric tests. This consideration is also the verification of test assumptions. If the permutation or the bootstrap is chosen to calibrate the significance test, a proper test statistic should be selected, consistent with how the original hypotheses are formulated.

The issue with large sample sizes also lies in choosing a significance test. If the sample size is extremely large, then tests based on resampling methods such as the permutation and the bootstrap may be computationally expensive. In such case, sampling-free tests are

preferred.

Again, choosing a test with the most significant result after using many tests commits p-hacking, which should be avoided by all means.

6.3 Transparent and complete reporting

After running a significance test on the data, usually a p-value or sometimes a confidence interval is obtained. Given a predetermined significance level α, the null hypothesis can be rejected if the p-value is smaller than α; otherwise, we fail to reject the null. In multiple comparison, the adjustment to the p-values is necessary before drawing the conclusion. If a confidence interval is obtained and it does not contain the value of test statistic specified in the null hypothesis (for example, mean difference is 0), then we can reject the null.

In reporting the test result, one should include the following results (use the comparison of means as an example): **the sample size**, **Cohen's** d (or other indices of effect size), **the observed test statistic**, **the** p**-value** and **the test used**.

In drawing a conclusion, one should refrain from only relying on the obtained p-value. Note that in case study 3, we have demonstrated that even though the test result suggests there exists statistically significant difference between the means of the two samples, there is a huge portion of overlapping data points in the histograms, and the effect size is trivial. The significance exists because of the extremely large sample size. Large sample sizes of test sets are not rare in NLP system performance comparison tasks. Therefore reporting the effect size is necessary, even if the test result is significant.

Instead of giving point estimates, reporting confidence intervals for the true test statistic and the true effect size if one wishes to quantify uncertainty is usually recommended. Also give summary statistics for reproduction of the results.

6.4 Interpretation of confidence intervals

Confidence intervals are alternative to point estimates of statistics of interest. A 95% confidence interval for the population mean implies that if we repeatedly sample from the popu-

lation and calculate the confidence interval, 95% of the intervals will contain the true population mean. It is misleading to think that the confidence interval contains the true mean with 0.95 probability. There exists duality between the confidence interval and a statistical hypothesis test: a two-sided confidence interval corresponds to a two-sided test; a one-sided confidence interval corresponds to a one-sided test. Suppose we would like to compare the mean difference with 0, suggesting a null hypothesis that the true mean difference is 0. Conventionally, the point 0 lying outside the confidence interval is equivalent to a rejection of the null hypothesis.

6.5 *i.i.d.* in categorical data

The major remaining obstacle lies in the violation of *i.i.d.* assumption for many tests, given the dependent nature of natural language data. Although in case study 3, we partially attempt to address this problem using the block bootstrap, we are still assuming independence among articles, which is yet to be verified further. In case study 4 of the two-sample categorical data, McNemar's test is used but note that each observation in the two samples might be dependent on its neighbors, violating the assumption of random sampling, since a sample due to random sampling must contain independent observations. Such dependency can be seen from the fact that a token in the center of a chunk, which is predicted with I (inside), is likely to have their neighbors predicted with I as well. Also, the probability that a B tag follows an O tag is 1. Therefore, we should find an alternative test for paired and dependent categorical samples to ensure the dependency structure is respected. In our work, we ignored the dependency and proceeded as if the observations are independent.

6.6 Test statistic for block bootstrap

In case study 3, we used the mean difference as the test statistic and computed the confidence interval based on the t ratios. The t ratios were calculated using the standard error of the entire sample. Computing the standard error this way fails to respect the dependency among the observations in the sample. Hence, it is necessary to find a more appropriate test statistic

for the block bootstrap.

6.7 Power estimation for block bootstrap

In the power estimation of the block bootstrap in the third case study, the actual resampling procedure is done with respect to the indices of test instances in order to preserve pairing. Note that we do not sample the document id numbers (each sentence belongs to a document, which has an id number given in the test set) but instead individual test instances (sentences). The reason for this resampling approach is that each document has different numbers of sentences, which range from 10 to 100. Resampling the entire document will soon make the size of the bootstrapped sample too big. During the resampling procedure, the document id of which the resampled test instance is recorded, and the resampled sentences which originally belong to the same document will be considered a bootstrapped block of test instances. Recall that in the block bootstrap, a block or a cluster is a document consisting of multiple test instances, and the method resamples blocks with replacement.

This resampling approach, however, is innately flawed in a sense that it does not respect the original sampling procedure of the test set, which is supposed to sample a document in its entirety instead of sampling each individual test instance. In the original sampling procedure, if a particular sentence appears in the test set, then all other sentences in the same document will surely appear in the test set as well, which amounts to the fact that the test set does not contain independently and randomly sampled test sentences. Hence, it may not make sense to juxtapose the block bootstrap with other applicable tests in the power comparison in the third case study, and an alternative resampling procedure is required to assess the statistical power of the block bootstrap method.

To resampling in the manner given in this thesis, the bootstrapped sample then is in essence a simple random sample, consisting of independently chosen test instances from the original test set. Thus, the independence assumption made by the aforementioned significance tests (paired student t, Wilcoxon and sign tests) is somehow satisfied in the bootstrapped samples (if we pretend not to have a priori knowledge of the test instances).

Therefore, there is no way to estimate the true power curves of these three tests which are thought to have no applicability to this case where dependency within the test set exists. If we instead sample with replacement the documents as a whole, then the sample size will grow fast, and we will again be confronted with the issues of over-powerfulness, which will bring significant results regardless of the true effect size.

6.8 Cross-domain comparison

This thesis only considers the system performance comparison tasks using a single test set within the same domain. For example, a test set for machine translation system performance comparison consists of only news. Thus, it is safe to assume that the sample contains identically distributed observations. Nowadays to build more robust NLP systems, the technique of domain adaptation is integrated into those systems. Thus, evaluating the systems with domain adaptation using a single-domain test set is not sufficient, necessitating cross-domain comparison of system performance over multiple test sets from different domains. This issue is addressed in Demšar (2006), which explored performance comparison tasks of classifiers over multiple test sets, whereas the test sets may not be cross-domain.

In cross-domain test sets, essentially we are confronted with the issue that samples contain non-identically distributed observations (potentially in clusters or blocks). When we have random samples of non-homogeneously distributed observations, one way to avoid the assumption of identical distribution is to use tests that only rely on independence among observations. Many nonparametric tests do not rely on such assumption but only assume independence. For example, the Wilcoxon signed-rank test, the Mann-Whitney U test and the sign test only rely on the independence assumption.

6.9 Choice of evaluation metric

This thesis did not address the issue related to choosing an evaluation metric for a particular NLP task. The statistical significance testing can also play a role in choosing an evaluation metric for a task. Consider choosing a metric for a machine translation evaluation task.

Traditionally the quality of translation depends on human assessment. To take human assessment as the gold standard, we can determine how the proposed evaluation metric is positively correlated with the human judgment. In this case, a test for correlation will become necessary.

Chapter 7
CONCLUSION AND FUTURE WORK

In this section, we summarize and highlight contributions of this thesis and provide directions for prospective studies on the significance testing in NLP.

7.1 Contribution

This thesis has accomplished the following:

1. We explored the use of the statistical significance testing in the system performance comparison in NLP by formally formulating the system performance task under the statistical significance testing framework.

2. We introduced in great detail a selection of relevant statistical significance tests including parametric tests, nonparametric tests and tests based on randomization techniques (bootstrap and permutation), with their corresponding test assumptions, testing procedures and their practical and potential usage in NLP explicated.

3. Special attention has been devoted to the way and the importance of verifying test assumptions, the choice of an appropriate test under different scenarios, the proper use of the chosen test and the way to draw a statistically sound conclusion based on the test result.

4. We advocated the reporting of the test result in a more transparent and complete fashion for the sake of replicability and reliability. To achieve this goal, the notion of effect size and power analysis were introduced and explained.

5. Four case studies with simulated and real data were presented as illustrative examples to demonstrate testing procedures. Each case study corresponded to a frequently seen data type in NLP system performance comparison tasks, where we first proposed a hypothesis of interest, simulated or gathered data, selected an appropriate test based on whether the test assumptions were met, applied the test, reported the test result in full transparency and detail and conducted post hoc exploratory analysis and power analysis to examine how the test performed.

6. The block bootstrap was introduced and employed in testing of performance of two machine translation systems due to special consideration of potential dependency within the test set.

7. In addition, we discussed comparisons with previous studies, remaining issues and limitations of this thesis from several perspectives to guide future researches.

7.2 Future trajectories

This thesis only discusses the use of the statistical significance testing in the scenario of system performance comparison. The statistical significance testing, as noted before, can be applied to a wider range of NLP problems. Also, this thesis only considers system performance comparison based on a single test set of homogeneous test instances. As the demand for more robust and domain-independent NLP systems arises, significance testing on cross-domain test sets poses a new statistical challenge since the vital assumption of *i.i.d.* is violated. We propose that prospective studies could be conducted in the following trajectories:

1. Apply or develop new tests for non-*i.i.d.* categorical data.

2. Instead of point estimates, use the bootstrap method to provide more accurate confidence intervals for a test statistic of interest such as the mean difference and the

median difference of evaluation metric values of two systems; also consider using a more appropriate test statistic for the block bootstrap.

3. To build and evaluate a more robust NLP system, a test set of cross-domain observations will be needed in the evaluation process. Testing based on cross-domain test set brings issues with non-*i.i.d.* observations within the test set. Thus it is necessary to develop an appropriate multiple testing procedure for cross-domain comparison.

4. The impact of the different choices of evaluation metrics for the same NLP task on the significance testing in system performance comparison remains to be explored.

5. As a more general topic, meta-analysis as well as different types of power analysis in NLP system performance comparison is to be discussed.

Appendix A

TABLES

A.1 Tables of R and Python functions for statistical tests

Table A.1 lists *R* functions of a selection of frequently used statistical tests. Arguments given here are not comprehensive. For a complete list of function arguments, use *RStudio* and type the question mark ? followed by the function name. For both chi-square tests and McNemar's test, the argument `x` can be a 2×2 contingency table or a vector, the latter of which requires `y` (a factor denoting the level of each observation in `x`) be specified. Note that the Wilcoxon signed-rank test and the Mann-Whitney U test use the same *R* function, with different `paired` arguments. The argument `x` can be a vector without specifying `y` if a one-sample testing is desired. If a two-sample test is of interest, then specify `y`. In the Friedman test, the argument `y` can be a matrix of data of a vector, the latter of which requires arguments `groups` and `blocks` be specified.

Table A.2 lists *Python* functions of a selection of frequently used statistical tests. Note that the F test is not implemented in *Python* (functions for ANOVA exist). The arguments for the Friedman test are array-like.

The bootstrap test, block bootstrap, the sign test and the permutation test are not listed in this table. These tests based on randomization techniques are customized *R* functions[1].

[1]Please see `https://github.com/haz060/NLPsigtest/actions` for more *R* code.

Table A.1: The table of R functions for tests

Test name	**Function**	**Package**
t test	`t.test(x, y, alternative, mu=0, paired=`**`FALSE`**`, var.equal=FALSE)`	`stats`
Paired t test	`t.test(x, y, alternative, mu=0, paired=`**`TRUE`**`, var.equal=FALSE, ...)`	`stats`
F test	`var.test(x, y, alternative, ...)`	`stats`
χ^2 GOF test	`chisq.test(x, y=NULL, ...)`	`stats`
χ^2 independence test	`chisq.test(x, y=NULL, ...)`	`stats`
McNemar's test	`mcnemar.test(x, y=NULL, correct=TRUE)`	`stats`
Fisher's exact test	`fisher.test(x, y=NULL, ...)`	`stats`
Wilcoxon signed-rank test	`wilcox.test(x, y=NULL, alternative, mu=0, paired=`**`TRUE`**`, ...)`	`stats`
Mann-Whitney U test	`wilcox.test(x, y=NULL, alternative, mu=0, paired=`**`FALSE`**`, ...)`	`stats`
Friedman test	`friedman.test(y, groups=NULL, blocks=NULL)`	`stats`
Shapiro-Wilks test	`shapiro.test(x)`	`stats`
Pearson normality test	`pearson.test(x, ...)`	`stats`
2-sample Kolmogorov-Smirnov test	`ks.test(x, y, alternative, exact = NULL)`	`stats`
Anderson-Darling test	`ad.test(x)`	`stats`

Table A.2: The table of *Python* functions for tests

Test name	**Function**	**Package**
t test	`scipy.stats.ttest_ind(a, b, equal_var=`**`True`**`, ...)`	`scipy.stats`
Paired t test	`scipy.stats.ttest_rel(a, b, ...)`	`scipy.stats`
F test	Not implemented	None
χ^2 GOF test	`scipy.stats.chisquare(f_obs, f_exp=None, ...)`	`scipy.stats`
χ^2 independence test	`scipy.stats.chi2_contingency(observed, ...`	`scipy.stats`
McNemar's test	`statsmodels.stats.contingency_tables.mcnemar(table, ...)`	`statsmodels.api`
Fisher's exact test	`scipy.stats.fisher_exact(table, alternative)`	`scipy.stats`
Wilcoxon signed-rank test	`scipy.stats.wilcoxon(x, y=None, alternative, ...`	`scipy.stats`
Mann-Whitney U test	`scipy.stats.mannwhitneyu(x, y, alternative=None, ...)`	`scipy.stats`
Friedman test	`scipy.stats.friedmanchisquare(...)`	`scipy.stats`
Shapiro-Wilks test	`scipy.stats.shapiro(x)`	`scipy.stats`
Pearson normality test	`scipy.stats.normaltest(a, ...`	`scipy.stats`
2-sample Kolmogorov-Smirnov test	`scipy.stats.ks_2samp(data1, data2)`	`scipy.stats`
Anderson-Darling test	`scipy.stats.anderson(x, dist)`	`scipy.stats`

A.2 Table of R functions for effect size indices

This section contains the table of R functions to compute some indices of effect size. Note that we do not include corresponding *Python* because there is no package that is specifically dedicated to power analysis and effect size in *Python*. The computation of these following indices can be done manually with knowledge of summary statistics such as the sample mean and variance, which can be obtained using *Python*.

Table A.3: The table of tests

Name	**Function**	**Package**
Cohen's d	`cohen.d(x, y, paired, pooled, mu, hedges.correction=FALSE,...)`	`effsize`
Hedges' g	`cohen.d(x, y, paired, pooled, mu, hedges.correction=TRUE,...)`	`effsize`
Cohen's g	`cohenG(x, digits)`	`rcompanion`
Odds ratio	`cohenG(x, digits)`	`rcompanion`

Appendix B

ALGORITHMS

B.1 Algorithm for unpaired permutation test

Algorithm 1 Unpaired permutation Test

Input: X, Y, two independent samples

Output: p, permutation p-value

procedure POOLED PERMUTATION
    Calculate the observed test statistic of choice $\theta_{ob}(X, Y)$
    Given a large number B
    Given a counting number $C \leftarrow 0$
    Concatenate X and Y, named Z
    for $b \in 1 : B$ **do**
        Shuffle Z and divide into X_p, Y_p
        Calculate permutation test statistic $\theta_p(X_p, Y_p)$
        if $\theta_p \geq \theta_{ob}$ **then**
            $C \leftarrow C + 1$
    Calculate permutation p value $= \frac{C+1}{B+1}$
    return p-value

B.2 Algorithm for paired sign test

Algorithm 2 Paired permutation sign 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

B.3 Algorithm for unpaired bootstrap test

Algorithm 3 Unpaired bootstrap Test

Input: X, Y, two independent samples

Output: p, bootstrap p-value

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

B.4 Algorithm for block bootstrap test

Algorithm 4 Block Bootstrap Test

Input: (X, Y), two paired samples; membership indicator $\gamma(\cdot, \cdot) \in \Gamma = \{0, 1, \cdots\}$

Output: p, permutation p-value

procedure Block Bootstrap
    Calculate the observed test statistic of choice $\theta_{ob}(X, Y)$
    Given a large number B
    Given a counting number $C \leftarrow 0$
    for $b \in 1 : B$ **do**
        Group (X, Y) into blocks where each observation has the same $\gamma(x_i, y_i)$.
        Independently sample blocks in (X, Y) with replacement and obtain (X_b, Y_b)
        Calculate bootstrap test statistic $\theta_b(X_b, Y_b)$
        if $\theta_b \geq \theta_{ob}$ **then**
            $C \leftarrow C + 1$
    Calculate bootstrap p value $= \frac{C+1}{B+1}$
    return p-value

B.5 Algorithm for Monte Carlo power estimation

Algorithm 5 MC power simulation

Input: $T(\cdot)$, a statistical test; α, sig. level

Output: p, power curve

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

B.6 Algorithm for bootstrap power estimation

In this algorithm, we use a double loop to obtain the power curves for different sample sizes. The main idea is to operate under the alternative hypothesis. To do so, we manually shift the data by the amount of the observed effect size (or an effect size that the user wish to have). The shifting is done in accordance with the alternative hypothesis. Suppose that the alternative is $\mu_X > \mu_Y$, we shift the sample X by δ.

Algorithm 6 Bootstrap power simulation

Input: X,Y, data; δ, effect size; $T(\cdot)$, a test; α, 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 δ depending on the alternative
            Run test $T(X^b, Y^b, \alpha)$
            Obtain a p-value
        $p(n) \leftarrow$ proportion$(\hat{p} < \alpha)$
    return p

BIBLIOGRAPHY

Amidei, Jacopo, Paul Piwek, and Alistair Willis (2019). "The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 397–402.

Benjamini, Y. and Y. Hochberg (1995). "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing". In: *J. Royal Statist. Soc., Series B* 57, pp. 289 –300.

Berg-Kirkpatrick, T., D. Burkett, and D. Klein (2012). "An Empirical Investigation of Statistical Significance in NLP". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 995–1005.

Bisani, M. and H. Ney (2004). "Bootstrap estimates for confidence intervals in ASR performance evaluation". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1, pp. I –409.

Blitzer, John, Ryan McDonald, and Fernando Pereira (2006). "Domain Adaptation with Structural Correspondence Learning". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia: Association for Computational Linguistics, pp. 120–128. URL: https://www.aclweb.org/anthology/W06-1615.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi (2017). "Findings of the 2017 Conference on Machine Translation (WMT17)". In: *Proceedings of*

the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Copenhagen, Denmark: Association for Computational Linguistics, pp. 169–214.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang (2007). "Word Sense Disambiguation Improves Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 33–40.

Chen, Zhiyuan, Arjun Mukherjee, and Bing Liu (2014). "Aspect Extraction with Automated Prior Knowledge Learning". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 347–358.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith (2011). "Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 176–181.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

— (1994). "The earth is round ($p < .05$)". In: *American Psychologist*, pp. 997–1003.

Collings, B. and M. Hamilton (1988). "Estimating the Power of the Two-sample Wilcoxon Test for Location Shift". In: *Biometrics* 44.3, pp. 847–860.

Collins, Michael, Philipp Koehn, and Ivona Kučerová (2005). "Clause Restructuring for Statistical Machine Translation". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 531–540.

Cooper, H., Larry Hedges, and Jeffrey Valentine (2009). *The Handbook of Research Synthesis and Meta-Analysis (Second Edition)*. Vol. 113, pp. 207–243.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Déjean, Hervé (2000). "Learning Syntactic Structures with XML". In: *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Demšar, Janez (2006). "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine Learning Research* 7, pp. 1–30.

Dror, R., G. Baumer, S. Shlomov, and R. Reichart (2018). "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1383–1392.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Faul, F., E. Erdfelder, A. Lang, and A. Buchner (2007). "GPower 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences". In: *Behavior Research Methods* 39, pp. 175–191.

Fisher, R. A. (1922). "On the Interpretation of χ^2 from Contingency tables, and the Calculation of P". In: *Journal of the Royal Statistical Society* 85.2, pp. 87–94.

Friedman, M. (1937). "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". In: *Journal of the American Statistical Association* 32.200, pp. 675–701.

Ghaeini, Reza, Sadid A. Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri (2018). "DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1460–1469.

Graham, Yvette, Nitika Mathur, and Timothy Baldwin (2014). "Randomized Significance Tests in Machine Translation". In: *Proceedings of the Ninth Workshop on Statistical Ma-*

chine Translation. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 266–274.

Halteren, Hans van (2000). "Chunking with WPDV Models". In: *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Jang, Hyeju, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rosé (2016). "Metaphor Detection with Topic Transition, Emotion and Cognition in Context". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 216–225.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, pp. 388–395.

Kumar, Shankar and William Byrne (2004). "Minimum Bayes-Risk Decoding for Statistical Machine Translation". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, pp. 169–176.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses*. Third. Springer Texts in Statistics. Springer.

Lin, Mingfeng, Henry Lucas, and Galit Shmueli (2013). "Too Big to Fail: Large Samples and the p-Value Problem". In: *Information Systems Research* 24, pp. 906–917.

Mann, H. B. and D. R. Whitney (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *Ann. Math. Statist.* 18.1, pp. 50–60.

McGraw, Kenneth O. and S. P. Wong (1992). "A common language effect size statistic". In: *American Psychological Association* 111(2), pp. 361–365.

McNemar, Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157.

Pearson, F.R.S. K. (1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302, pp. 157–175.

Perugini, Marco, Marcello Gallucci, and Giulio Costantini (2018). "A Practical Primer To Power Analysis for Simple Experimental Designs". In: *International Review of Social Psychology* 31.

Rush, Alexander, Roi Reichart, Michael Collins, and Amir Globerson (2012). "Improved Parsing and POS Tagging Using Inter-Sentence Consistency Constraints". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1434–1444.

Sadat, Fatiha, Masatoshi Yoshikawa, and Shunsuke Uemura (2003). "Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach". In: *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*. Sapporo, Japan: Association for Computational Linguistics, pp. 57–64.

Shapiro, S. S. and M. B. Wilk (1965). "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3-4, pp. 591–611.

Smucker, M., J. Allan, and B. Carterette (2007). "A comparison of statistical significance tests for information retrieval evaluation". In: *International Conference on Information and Knowledge Management, Proceedings*, pp. 623–632.

Snedecor, G. W. (1934). "Calculation and Interpretation of Analysis of Variance and Covariance". In: *Iowa State College Division of Industrial Science monographs* 1.

Søgaard, Anders (2013). "Estimating effect size across datasets". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 607–611.

Søgaard, Anders, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso (2014). "What's in a p-value in NLP?" In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 1–10.

Štefanec, Vanja, Nikola Ljubešić, and Jelena Kuvač Kraljević (2016). "Croatian Error-Annotated Corpus of Non-Professional Written Language". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia, pp. 3220–3226.

Student (1908). "The Probable Error of A Mean". In: *Biometrika* 6.1, pp. 1–25.

Trinh, Anh Duong, Robert J. Ross, and John D. Kelleher (2019). "Capturing Dialogue State Variable Dependencies with an Energy-based Neural Dialogue State Tracker". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, pp. 75–84.

Wasserstein, R. L. and N. A. Lazar (2016). "The ASA Statement on p-Values: Context, Process, and Purpose". In: *The American Statistician* 70.2, pp. 129–133.

Weeber, Marc, Rein Vos, and R. Harald Baayen (2000). "Extracting the lowest-frequency words: pitfalls and possibilities". In: *Computational Linguistics* 26.3, pp. 301–318.

Wilcoxon, F. (1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6, pp. 80–83.

Yang, Bishan and Tom Mitchell (2017). "Leveraging Knowledge Bases in LSTMs for Improving Machine Reading". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1436–1446.

Yang, Mei and Jing Zheng (2009). "Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, pp. 237–240.

Yeh, Alexander (2000). "More accurate tests for the statistical significance of result differences". In: *The 18th International Conference on Computational Linguistics*. Vol. 2.

Zumbo, Bruno D. and Anita M. Hubley (1998). "A Note on Misconceptions Concerning Prospective and Retrospective Power". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.2, pp. 385–388.