GOODNESS-OF-FIT TESTS UNDER PERMUTATIONS

by

Chen Chen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2019

Approved by:

Dr. Zhiyi Zhang

Dr. Jiancheng Jiang

Dr. Jun Song

Dr. Weidong Tian

ABSTRACT

CHEN CHEN. Goodness-of-fit Tests under Permutations. (Under the direction of DR. ZHIYI ZHANG)

Several new goodness-of-fit tests are proposed on countable alphabets, where certain fundamental statistical concepts associated with random variables, such as cumulative distribution functions, characteristic functions and moments, may not exist. An entropic perspective by ways of the entropic basis, derived from the well-known Turing's formula, is introduced. A new characterization theory of probability distributions on alphabets is established by means of the entropic basis. Based on this logic framework several goodness-of-fit tests are developed.

Toward developing the new goodness-of-fit tests, an one-to-one correspondence between a given probability distribution and its entropic basis is first established. In case the cardinality of underlying distribution is finite, say K, the first K entropic moments uniquely determine the underlying probability distribution up to a permutation on the index set. For each of the entropic moments, an uniformly minimum variance unbiased estimator (UMVUE) is introduced. Based on the sampling distribution of the UMVUEs of the entropic moments and the multivariate delta method, two new Chi-squared goodness-of-fit tests are constructed and their asymptotic distributional properties are established in theory. However it is also observed that these new tests are difficult to implement numerically. To alleviate the computational difficulty in implementation, a heuristic exact test for goodness-of-fit is proposed. The performance of the proposed tests are evaluated by simulation studies under a range of distributions. The new tests are also illustrated in several real life applications.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Zhiyi Zhang for giving me a chance to study statistics, passing on his knowledge and training me to be a good statistician. I would also like to thank my committee members, Dr. Jiancheng Jiang, Dr. Jun Song and Dr. Weidong Tian.

I would like to thank all the professors and colleagues who gave me instructions and inspirations. I would also like to thank the Department of Mathematics and Statistics and the Graduate School at UNC Charlotte for the financial support.

I would like to thank my parents, for their unconditional love and warm support. I would also like to thank my other family members, and friends. I'm blessed having them in my life.

I can not imagine how I would accomplish this without any of these people.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

1.1 Alphabet and Goodness-of-fit Test

Let $\mathscr{X} = \{\ell_k; k \geq 1\}$ be a countable and categorical sample space, where each category is assigned with a label ℓ_k for some k. In many information theory literatures, this kind of sample space is referred to as an *alphabet*, and those labels are called *letters* [1]. Let $\mathbf{p} = \{p_k; k \geq 1\}$ where $p_k > 0$ for every k, be a probability distribution associated with $\mathscr{X}$. Let $\mathbf{S} = \{X_i; i = 1, \cdots, n\}$ be an identically and independently distributed (*i.i.d.*) sample of size n drawn from $\mathscr{X}$ under $\mathbf{p}$. Let the sample data be summarized into frequencies $\mathbf{Y} = \{Y_k; k \geq 1\}$ and relative frequencies $\hat{\mathbf{p}} = \{\hat{p}_k = \frac{Y_k}{n}; k \geq 1\}$.

By using the name *alphabet*, as opposed to the usual sample space where random variables reside, we emphasize that, under the consideration of this dissertation, no metric is required nor imposed. All letters don't have to be numeric, not even ordinal, and can be purely nominal like "labels". The central concept of modern probability theory and statistics is random variable, which is a measurable function that maps the sample space into a real space. But what if the sample space can not be metricized properly? Then a random variable can not be well-defined, consequently many usual statistical concepts such as cumulative distribution function, characteristic function, and moments no longer exist. This issue is becoming more and more common in modern data science, since we're facing many challenges from high dimensionality, uncertain data type, complex data structure, and so on. Theoretically, one may manually assign a numeric order to letters or use dummy variables, but that doesn't make too much sense for interpretability. For example, people love to use L_2 metric since we live in a 3-dimensional space and are very familiar with Euclidean distance.

However, when dimensionality increases to a million or billion level, the meanings of many common quantitative concepts including Euclidean distance become unclear, or non-existent, *aka. curse of dimensionality* [2]. For another example, when dealing with qualitative data like personal names, it is almost impractical to find a totally prescribed and meaningful numeric metric. However, probabilities, or proportions of all letters can always be defined, without using any metric or concepts like random variable. We define a variable that randomly takes values on an alphabet as a *random element.* The collection of all possible values of a random element, together with the probabilities for each value, is called a probability distribution on the alphabet.

When given a distribution and a random sample, a very basic objective in statistics is to check that, does this sample come from this distribution [3]. A hypothesis test that asserts whether a given distribution is suited to a sample, is called goodness-of-fit test, who plays an important role in many areas like data mining and model validation. The problem can be stated in formal mathematical language as follows. Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a probability distribution on $\mathscr{X}$, with p_ks all unknown. Let $\mathbf{S} = \{X_i; i = 1, \cdots, n\}$ be an *i.i.d.* sample of size n drawn from $\mathscr{X}$ under $\mathbf{p}$. Let $\mathbf{q} = \{q_k; k \geq 1\}$ be another probability distribution on $\mathscr{X}$, with q_ks all pre-specified and known. The goodness-of-fit test is to check the hypothesis:

$$H_0 : \mathbf{p} = \mathbf{q} \qquad vs. \qquad H_a : \mathbf{p} \neq \mathbf{q} \tag{1.1}$$

based on sample data $\mathbf{S}$.

1.2 Issues of Classical Tests

In classical statistics, many goodness-of-fit tests have been proposed, for example, Kolmogorov-Smirnov test [4], Anderson-Darling test [5], Pearson's Chi-squared test [6], Multinomial test [3], G-test [7], etc. But sometimes, it's quite different, and challenging to test goodness-of-fit on alphabets, mainly due to two issues.

The first issue is about metric. Let's look at the statistic of Kolmogorov-Smirnov test,

$$D = \sup_x |F_n(x) - F(x)| \tag{1.2}$$

and that of Anderson-Darling test,

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)\,(1 - F(x))}\, dF(x) \tag{1.3}$$

where $F(x)$ is the cumulative distribution function of underlying distribution $\mathbf{q}$, and $F_n(x)$ is the empirical distribution function of given sample $\mathbf{S}$ of size n. It is clear those two statistics wholly rely on cumulative distribution functions. However, as we stated above, a numeric order as well as a valid cumulative distribution function can not be guaranteed on alphabets. As a result, those tests based on CDF may not work.

The second issue is about *linkage*, or more precisely, the linkage between a random sample and the underlying distribution. Again, let us look at the statistic of Pearson's Chi-squared test,

$$\chi^2 = \sum_{k=1}^{K} \frac{\left(Y_k - E_k\right)^2}{E_k} \tag{1.4}$$

and that of G-test,

$$G = 2 \sum_{k=1}^{K} Y_k \cdot \ln\left(\frac{Y_k}{E_k}\right) \tag{1.5}$$

where $E_k = n \cdot q_k$ is the expected frequency for letter ℓ_k under distribution $\mathbf{q}$. As we can see, those two tests work only when each pair of Y_k and E_k is one-to-one matched, *ie.*, they need a perfect pairwise linkage between the sample and underlying distribution. If we don't have enough information about this linkage, then all such kind of tests no longer work. One may wonder does this issue really happen in practice? The answer is yes. For your understanding, let's consider the following two scenarios.

Scenario 1. The linkage is well-defined, but our sample data set is incomplete or damaged. We only have a set of frequencies that counting from different letters, but we don't clearly know which frequency is for which letter.

Scenario 2. The linkage can not be pre-specified before the random experiment. For a simple example, consider drawing n chips from a box containing chips of K different colors, distinguishable but unspecified colors. In this case, an assignment of which k is which color is not possible, and is not necessary for the experiment to be carried out and data collected.

In classical statistics, before an experiment is conducted, the sample space is often completely prescribed, that is, every possible outcome of the experiment is completely describable and identifiable when observed. This specificity of sample space is relaxed in different ways and to different degrees in some situations of modern data science, partially inspired by the empirical Bayesian school of thought and partially due to the data complexity and high dimensionality. The said specificity, or the lack of it, could vary over a wide spectrum. To put this argument in a broader perspective, one may view many statistical problems in modern data science as those with countable discrete sample spaces, non-metricized, non-ordinal, not completely prescribed (*ie., alphabets*), but with distinguishable elements (*ie., letters*). This is another important reason why we introduce alphabet and letter concepts at the beginning, rather than using usual sample space settings.

In summary, the metric issue and linkage issue on alphabets challenge traditional goodness-of-fit tests, and we need to find some new tests.

1.3 A Weaker Hypothesis

Let's consider an alternative hypothesis:

$$H_0 : \mathbf{p}_\downarrow = \mathbf{q}_\downarrow \qquad vs. \qquad H_a : \mathbf{p}_\downarrow \neq \mathbf{q}_\downarrow \tag{1.6}$$

where the sub-index $\downarrow$ denotes a decreasingly ordered probability distribution, *ie.*, $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$, where $p_{(1)}$ is the maximum of all p_ks, $p_{(2)}$ is the second largest, so and so on. Similar notations are also defined for $\mathbf{q}_\downarrow$. Noting that $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ is a weaker statement than $\mathbf{p} = \mathbf{q}$, in the sense that the latter implies the former but not vice versa. This can be also viewed as a generalized hypothesis, in the sense that we no longer focus on a single distribution $\mathbf{p}$, but on a family of different distributions, all of which share the same invariant $\mathbf{p}_\downarrow$ under permutations.

An instant benefit of the hypothesis in (1.6) is that, it doesn't suffer the metric issue and linkage issue. Since all p_ks, as probabilities, are real numbers between 0 and 1, they can always be well and easily ordered, no matter whether the letters are ordinal or not, whether the linkage information is available or not.

The utility of the hypothesis in (1.6) is seen more readily in an alternative form. Consider the family of all functionals, denoted $\mathscr{F}$, such that each of its members, denoted F, satisfies $F(\mathbf{p}) = F(\mathbf{q})$ if and only if $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$. The hypothesis of (1.6) can then be equivalently represented by

$$\begin{aligned} &H_0 : F(\mathbf{p}_\downarrow) = F(\mathbf{q}_\downarrow) \text{ for all } F \in \mathscr{F} \qquad vs. \\ &H_a : F(\mathbf{p}_\downarrow) \neq F(\mathbf{q}_\downarrow) \text{ for some } F \in \mathscr{F} \end{aligned} \tag{1.7}$$

In modern data science, the energy in a random data field is often summarized by functionals of $\mathbf{p}_\downarrow$ that are invariant under permutations on on the index set $\{k; k \geq 1\}$. For example, in information theory, many types of information are summarized by functionals such as Shannon's entropy [8] or mutual information [9]; and in ecology, the concept of diversity (index) is often measured by functionals such as Rényi's entropy or Simpson's index. To see a list of such indices, one may refer to Zhang and Grabchak (2016) [10]. Each particular functional represents a particular perspective to an underlying interest, which varies from situation to situation. The family $\mathscr{F}$

represents the totality of all such functionals. A non-rejection of H_0 in (1.7) indicates a lack of evidence for a shift with any $F \in \mathscr{F}$, while a rejection would encourage further research into identifying finer features of the difference between $\mathbf{p}$ and $\mathbf{q}$. The hypothesis in (1.7) is a general hypothesis, paralleling the logic structure of the F-test in detecting differences among multiple treatment effects in a classical ANOVA setting.

It's worth to mention that, we propose several new goodness-of-fit tests mainly to overcome the metric issue and linkage issue on alphabets, but those new tests work for numerical variable and sample carrying on linkage information as well. More interesting, the new tests perform even better than traditional tests when they both work, especially when sample size n is relatively small, compared to distribution cardinality K.

This dissertation is organized as follows. In Chapter 2, we introduce the entropic moments, prove the one-to-one correspondence between entropic basis and underlying distribution, and give sampling distribution of entropic moments. In Chapter 3, we present the central results of this dissertation, including the construction of two new Chi-squared goodness-of-fit tests and a heuristic test, together with theoretical analysis and simulation studies. In Chapter 4, two real data examples are demonstrated. In Chapter 5, we introduce an R package "Entropic", which provides core functions to implement entropic perspective related computations. Some detailed simulation results, additional data and descriptions are provided in Appendix.

CHAPTER 2: Entropic Perspective on Alphabets

Let $\mathbf{p}$ be a probability distribution as defined in Chapter 1. For each $\mathbf{p}$ and any positive integer u, let

$$\eta_u = \eta_u(\mathbf{p}) = \sum_{k \geq 1} p_k^u \tag{2.1}$$

be referred to as the u^{th} entropic moment [10].

2.1 A New Characterization

Zhang and Zhou (2010) [11] gave this result:

Lemma 1. *Let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions on the same countable alphabet $\mathscr{X}$. Then $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ if and only if $\eta_u(\mathbf{p}) = \eta_u(\mathbf{q})$ for all integers $u \geq 1$.*

It is stated that for any probability distribution $\mathbf{p} = \{p_k; k \geq 1\}$, including those with countably infinite $p_k > 0$, $\{\eta_u; u \geq 1\}$ uniquely determines $\mathbf{p}_\downarrow$. In fact, it can be shown that any tail of the infinite sequence $\{\eta_u; u \geq u_0\}$ for any fixed $u_0 \geq 1$, uniquely determines $\mathbf{p}_\downarrow$. Now we see all entropic moments together can work as a characterization of probability distributions on alphabets.

Further, when the cardinality K of probability distribution is finite, we have a stronger result, as stated in the following theorem. Let $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{p}) = \{\eta_u; u = 1, \cdots, K\}$ be referred to as the entropic basis.

Theorem 1. *Let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions on the same finite alphabet $\mathscr{X}$. Then $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ if and only if $\boldsymbol{\eta}(\mathbf{p}) = \boldsymbol{\eta}(\mathbf{q})$.*

Theorem 1 says that the first K entropic moments uniquely determine $\mathbf{p}_\downarrow$. Consequently justifies the following hypothesis as an equivalent form of (1.6).

$$H_0 : \boldsymbol{\eta}(\mathbf{p}) = \boldsymbol{\eta}(\mathbf{q}) \qquad vs. \qquad H_a : \boldsymbol{\eta}(\mathbf{p}) \neq \boldsymbol{\eta}(\mathbf{q}) \tag{2.2}$$

$\eta_u = \eta(u) = \sum_{k\geq 1} p_k^u$ may be viewed as a characteristic function or a moment generating function, in the sense that $\boldsymbol{\eta}(\mathbf{p})$ is obtained by being evaluated at positive integer values of u. $\boldsymbol{\eta}(\mathbf{p})$ may also be viewed as a re-parametrization of $\mathbf{p}_{\downarrow}$, and the re-parametrization has fundamental implications beyond the scope of this article. Interested readers may refer to Zhang (2018) [12] and Molchanov, Zhang and Zheng (2018) [13] for additional details.

As stated in Section 1.2 (Scenario 2), when the sample space indexes are not prespecified, the mere notion of $\mathbf{p} = \{p_k; k \geq 1\}$ is not well-defined, but $\mathbf{p}_{\downarrow}$ is and hence a legitimate object for inference. For that same reason, in modern data science, functionals of $\mathbf{p}_{\downarrow}$, *ie.*, $F \in \mathscr{F}$, are often of interest. For example, Shannon defined self-information to be associated with a distinguishable event ℓ_k as $-\ln p_k$, an information quantity not associated with the description (numerical or otherwise) of the event itself but only of its probability. Furthermore, on such sample spaces, the usual notions of moments, real or complex, are non-existent and therefore many classic theories of probability and statistics are no longer useful. However the notion of entropic moments provides a new characterization of the underlying $\mathbf{p}_{\downarrow}$. In short, the entropic basis, $\{\eta_u; u \geq 1\}$, has theoretical implications in its own right.

The complete proof of Theorem 1 can be found in Appendix A.1.

2.2 Sampling Distribution of Entropic Moments

The core support to the inferential procedure to be proposed in the subsequent text is the existence of an uniformly minimum variance unbiased estimator (umvue) of $\eta_u = \sum_{k \geq 1} p_k^u$ for every positive integer u, $u \leq n$ (sample size),

$$Z_u = \sum_{k \geq 1} \left[1_{[\hat{p}_k \geq u/n]} \prod_{j=0}^{u-1} \left(\frac{Y_k - j}{n - j} \right) \right] \tag{2.3}$$

proposed by Zhang and Zhou (2010) [11]. Noting $\eta_1 = Z_1 = 1$, let $\mathbf{Z}^* = (Z_2, Z_3, \cdots, Z_K)^\tau$ and $\boldsymbol{\eta}^* = (\eta_2, \cdots, \eta_K)^\tau$. The asymptotic distribution of entropic moments is derived in the following theorem.

Theorem 2. *For any given* $\mathbf{p} = \{p_k; k = 1, \cdots, K\}$ *satisfying* $p_k > 0$ *for each* k*,*

$$\sqrt{n}(\mathbf{Z}^* - \boldsymbol{\eta}^*) \xrightarrow{L} N(\mathbf{0}, \Sigma^*) \tag{2.4}$$

where $\mathbf{0}$ *is the* $(K-1)$*-dimensional column vector of zeros and* Σ^* *is a* $(K-1) \times (K-1)$ *covariance matrix as given in (2.5) below.*

Lemma 2 and Lemma 3 below are proposed to prove Theorem 2.

Lemma 2. *Let* $\hat{\eta}_u = \sum_{k=1}^K \hat{p}_k^u$ *for* $u = 1, \cdots, K$ *and* $\hat{\boldsymbol{\eta}}^* = (\hat{\eta}_2, \cdots, \hat{\eta}_K\}$*. Then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*) \xrightarrow{L} N(\mathbf{0}, \Sigma^*) \tag{2.5}$$

where $\Sigma^* = A^\tau \Sigma A$*,*

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_{K-1} \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_{K-1} & \cdots & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix} \tag{2.6}$$

and

$$A = \begin{pmatrix} 2(p_1 - p_K) & 2(p_2 - p_K) & \cdots & 2(p_{K-1} - p_K) \\ 3(p_1^2 - p_K^2) & 3(p_2^2 - p_K^2) & \cdots & 3(p_{K-1}^2 - p_K^2) \\ \vdots & \vdots & \ddots & \vdots \\ K(p_1^{K-1} - p_K^{K-1}) & \cdots & \cdots & K(p_{K-1}^{K-1} - p_K^{K-1}) \end{pmatrix} \tag{2.7}$$

Furthermore, Σ^* is of rank r, where r is such that $r+1$ is the number of distinct probabilities in $\mathbf{p}$.

For a given $\mathbf{p} = \{p_k; k \geq 1\}$, $\hat{\mathbf{p}} = \{\hat{p}_k = Y_k/n; k \geq 1\}$ from an *iid* sample of size n, and any positive integers $u \geq 1$ and $v \geq 1$, let

$$\zeta_{u,v} = \sum_{k=1}^{K} p_k^u (1 - p_k)^v \tag{2.8}$$

$$\hat{\zeta}_{1,v} = \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k)^v \tag{2.9}$$

$$Z_{1,v} = \sum_{k=1}^{K} \hat{p}_k \prod_{j=1}^{v} \left(1 - \frac{Y_k - 1}{n - j}\right) \tag{2.10}$$

Lemma 3. *For any $v \in \{0, 1, \ldots, K\}$, $n(Z_{1,v} - \hat{\zeta}_v) \xrightarrow{p} c$ as $n \to \infty$, where c is a constant.*

The complete proofs of Lemma 2, Lemma 3 and Theorem 1 are provided in Appendix A.2.

To summarize this chapter, entropic basis is a new characterization of probability distributions on alphabets, so inferences about distributions can be done through entropic moments and their estimators. The asymptotic normality in Theorem 2 and the availability of consistent estimators of Σ^* permit large sample confidence regions for the entropic moments $\boldsymbol{\eta}$, and hence a test for the hypothesis of (1.6) and (2.2), as will be described in next chapter.

CHAPTER 3: Hypothesis Testing

3.1 New Chi-squared Tests

Under the null hypothesis in (1.6), $H_0 : \mathbf{p}_\downarrow = \mathbf{q}_\downarrow = \{q_1, \cdots, q_K\}$, where $\mathbf{q}_\downarrow$ is completely specified, all the repeated values of q_k can be identified. Suppose there are $r+1$ distinct values of $\{q_k; k = 1, \cdots, K\}$. Denote these $r+1$ values and their multiplicities as

$q_{(1)}$	$q_{(2)}$	$\cdots$	$q_{(r)}$	$q_{(r+1)}$
m_1	m_2	$\cdots$	m_r	m_{r+1}

specifically noting that $q_{(1)} > q_{(2)} > \cdots > q(r+1) > 0$. For notation convenience, let $m_0 = m_{r+1} = 0$. Consequently $\mathbf{q}_\downarrow$ can be viewed as $r+1$ blocks of equal values, that is,

$$\mathbf{q}_\downarrow = \{q_{(1)}\mathbf{1}^\tau_{m_1}, q_{(2)}\mathbf{1}^\tau_{m_2}, \cdots, q_{(r)}\mathbf{1}^\tau_{m_r}, q_{(r+1)}\mathbf{1}^\tau_{m_{r+1}}\}$$

where $\mathbf{1}^\tau_m$ denotes a row vector of m "1"s. Consider a $r \times (K-1)$ matrix, C, of which the i^{th} row, of size $K-1$, consists of a sub-row of "$1/m_i$"s and of length m_i, and two other all-zero vectors of lengths $\sum_{j=0}^{i-1} m_j$ and $\sum_{j=i+1}^{r} m_j$ respectively. Letting $\mathbf{0}^\tau_m$ be a row of m "0"s,

$$C = \begin{pmatrix} \mathbf{0}^\tau_{m_0} & \frac{1}{m_1}\mathbf{1}^\tau_{m_1} & \mathbf{0}^\tau_{(K-1)-m_0-m_1} \\ \mathbf{0}^\tau_{m_1} & \frac{1}{m_2}\mathbf{1}^\tau_{m_2} & \mathbf{0}^\tau_{(K-1)-m_1-m_2} \\ \vdots & \vdots & \vdots \\ \mathbf{0}^\tau_{\sum_{j=0}^{i-1} m_j} & \frac{1}{m_i}\mathbf{1}^\tau_{m_i} & \mathbf{0}^\tau_{(K-1)-\sum_{j=0}^{i} m_j} \\ \vdots & \vdots & \vdots \\ \mathbf{0}^\tau_{(K-1)-m_r} & \frac{1}{m_r}\mathbf{1}^\tau_{m_r} & \mathbf{0}^\tau_{m_{r+1}} \end{pmatrix} \tag{3.1}$$

It can be verified that

$$AC^\tau = B \equiv \begin{pmatrix} 2(q_{(1)} - q_K) & 2(q_{(2)} - q_K) & \cdots & 2(q_{(r)} - q_K) \\ 3(q_{(1)}^2 - q_K^2) & 3(q_{(2)}^2 - q_K^2) & \cdots & 3(q_{(r)}^2 - q_K^2) \\ \vdots & \vdots & \ddots & \vdots \\ K(q_{(1)}^{K-1} - q_K^{K-1}) & K(q_{(2)}^{K-1} - q_K^{K-1}) & \cdots & K(q_{(r)}^{K-1} - q_K^{K-1}) \end{pmatrix}_{(K-1)\times r}$$

and therefore B is of full rank r because A_1 evaluated at $\mathbf{p} = \mathbf{q}_\downarrow$ is. This fact immediately gives the following corollary.

Corollary 1. *Under the null hypothesis $H_0 : \mathbf{p}_\downarrow = \mathbf{q}_\downarrow$, suppose that there are exactly $r+1$ distinct q_ks in $\mathbf{q}_\downarrow$ and that, $\boldsymbol{\eta}^* = (\eta_2, \cdots, \eta_K)^\tau$, Σ of (2.6), A of (2.7) and C of (3.1) are evaluated at $\mathbf{p} = \mathbf{q}_\downarrow$. Then*

1. $\sqrt{n}[C(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)] \xrightarrow{L} N(\mathbf{0}, CA^\tau\Sigma AC^\tau)$ *and* $CA^\tau\Sigma AC^\tau$ *is of full rank r; and*

2. $n[C(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)]^\tau(CA^\tau\Sigma AC^\tau)^{-1}[C(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)] \xrightarrow{L} \chi^2(r)$.

It is to be noted that if $\mathbf{q}_\downarrow$ is an uniform distribution then $r+1 = 1$, so A is of rank $r = 0$ and both limiting distributions of Corollary 1 degenerate; in fact, $r+1 = 1$ if and only if the underlying distribution is a uniform distribution. It may also be interesting to note that the action of C on $\hat{\boldsymbol{\eta}}^*$ as in $C(\hat{\boldsymbol{\eta}}^*)$ corresponds to taking averages in blocks of size m_i, $i = 1, \cdots, r$, in the $(K-1)$ dimensional vector $\hat{\boldsymbol{\eta}}^*$.

Theorem 2 gives another corollary as stated below.

Corollary 2. *Under the null hypothesis $H_0 : \mathbf{p}_\downarrow = \mathbf{q}_\downarrow$, suppose that there are exactly $r+1$ distinct q_ks in $\mathbf{q}_\downarrow$ and that, $\boldsymbol{\eta}^* = (\eta_2, \cdots, \eta_K)^\tau$, Σ of (2.6), A of (2.7) and C of (3.1) are evaluated at $\mathbf{p} = \mathbf{q}_\downarrow$. Then*

1. $\sqrt{n}[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)] \xrightarrow{L} N(\mathbf{0}, CA^\tau\Sigma AC^\tau)$ *and* $CA^\tau\Sigma AC^\tau$ *is of full rank r; and*

2. $n[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)]^\tau(CA^\tau\Sigma AC^\tau)^{-1}[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)] \xrightarrow{L} \chi^2(r)$.

3.1.1 Testing Procedures

Part (2) of Corollary 1 devices a large sample Chi-squared test for goodness-of-fit under permutations, which rejects H_0 if

$$T_p = n[C(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)]^\tau (CA^\tau \Sigma A C^\tau)^{-1} [C(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)] > \chi^2_\alpha(r) \tag{3.2}$$

where for some $\alpha \in (0,1)$, $\chi^2_\alpha(r)$ is the $100(1-\alpha)$ th percentile of the Chi-squared distribution with degrees of freedom r. This test is referred to in the subsequent text as the *plug-in test* T_p.

Part (2) of Corollary 2 devices another large sample Chi-squared test, which rejects H_0 if

$$T_z = n[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)]^\tau (CA^\tau \Sigma A C^\tau)^{-1} [C(\mathbf{Z}^* - \boldsymbol{\eta}^*)] > \chi^2_\alpha(r) \tag{3.3}$$

where for some $\alpha \in (0,1)$, $\chi^2_\alpha(r)$ is the $100(1-\alpha)$ th percentile of the Chi-squared distribution with degrees of freedom r. This test is referred to in the subsequent text as the *entropic test* T_z.

Two remarks may be made regarding the plugin test and the entropic test. First, in comparing Corollaries 1 and 2, the two tests are equally efficient asymptotically, noting specifically that the plugin test is based on the maximum likelihood principle. Second, one would expect the entropic test to perform better for finite samples since $\mathbf{Z}^*$ is an unbiased estimator of $\boldsymbol{\eta}^*$ but $\hat{\boldsymbol{\eta}}^*$ is not. In fact, Lemma 3 indicates that the decay rate of the bias of $\hat{\boldsymbol{\eta}}^*$ is much slower.

3.1.2 Simulations

In this section, we run numerical simulations to evaluate the plug-in test T_p and entropic test T_z, and compare them with the *linked* Pearson's Chi-squared test T_l.

As we stated in Section 1.2, the original Pearson's Chi-squared test doesn't work without linkage information. Here to use the numerical performance of Pearson's

Chi-squared test as a reference, we manually link each sample frequency to a letter in underlying distribution, based the numerical order. For example, the highest frequency in the sample will be linked to the largest probability in the underlying distribution, the second highest frequency will be linked to the second largest probability, and so on.

To evaluate size of those tests, we simply pick both the sampling and the underlying probability distribution to be:

$$\mathbf{p} = \left\{ \frac{5}{15}, \frac{4}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15} \right\} \tag{3.4}$$

as shown in Figure 3.1.

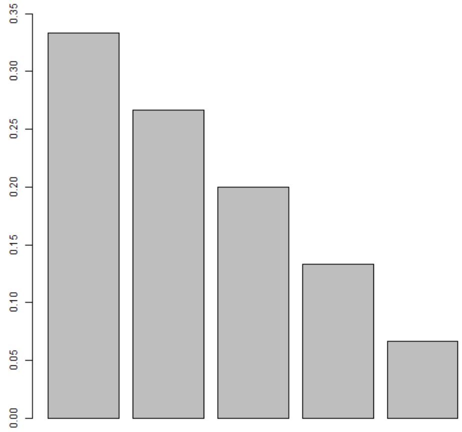

Figure 3.1: Underlying Distribution $\mathbf{p}$

And let $\alpha = 0.05$, number of iterations (for each sample size) $m = 100,000$, and sample sizes vary from 100 to $1000,000$. The simulation results are summarized into Table 3.1 and Figure 3.2.

As one can see, under the null hypothesis that a random sample is drawn from distribution $\mathbf{q}$, when the sample size is small ($n \leq 1,000$), the linked Pearson's test T_l tends to reject less than 0.05 of all random samples, while the plug-in test T_p

Table 3.1: Rejection Rates under H_0 with $\alpha = 0.05$

Sample Size	10^2	10^3	10^4	10^5	10^6
T_l	0.0083	0.0484	0.0502	0.0498	0.0506
T_p	0.2483	0.1101	0.0606	0.0505	0.0498
T_z	0.3677	0.1022	0.0571	0.0507	0.0501

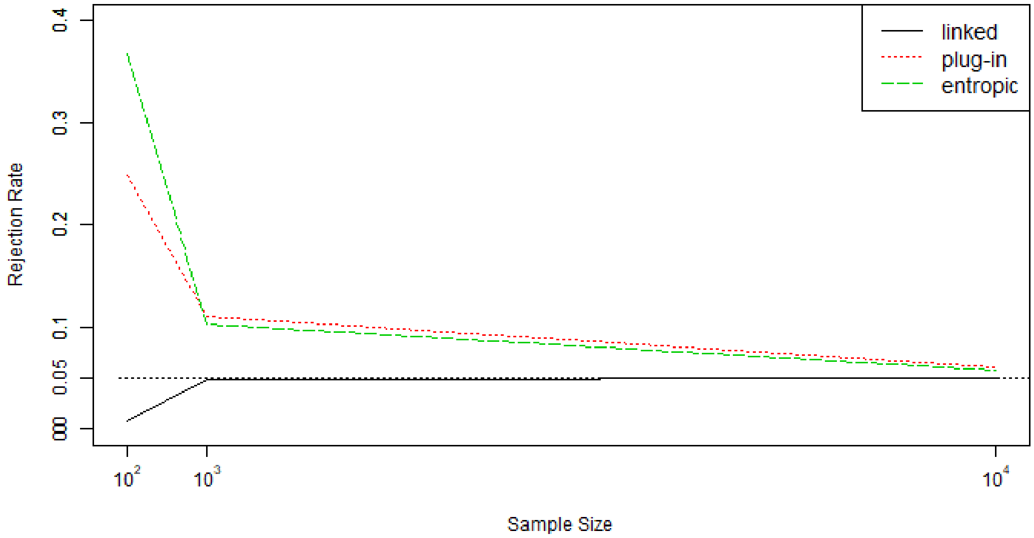

Figure 3.2: Rejection Rates under H_0 with $\alpha = 0.05$

and entropic test T_z tend to reject more than 0.05. When the sample size increases ($1,000 \leq n \leq 10,000$), size of T_l reaches 0.05 very fast, but T_p and T_z don't. When sample size is sufficiently large ($n \geq 10,000$), all 3 tests tend to have a size = 0.05 as expected.

This is not too surprising, because all 3 tests are derived from large sample distributions, and can't guarantee small sample performance.

Strictly speaking, if the test size cannot be controlled, then power analysis doesn't make too much sense. But for your reference and also to illustrate the differences between those 3 tests, we still run another simulation to examine test power.

Again, we pick the sampling probability distribution $\mathbf{p}$ as in (3.4), and pick the underlying probability distribution to be:

$$\mathbf{q} = \left\{ \frac{9}{35}, \frac{8}{35}, \frac{7}{35}, \frac{6}{35}, \frac{5}{35} \right\} \tag{3.5}$$

as shown in Figure 3.3 and Figure 3.4.

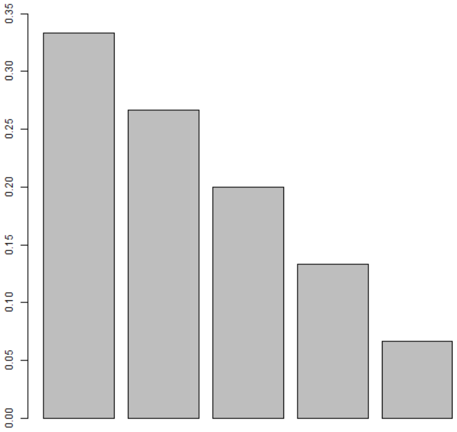

Figure 3.3: Sampling Distribution $\mathbf{p}$

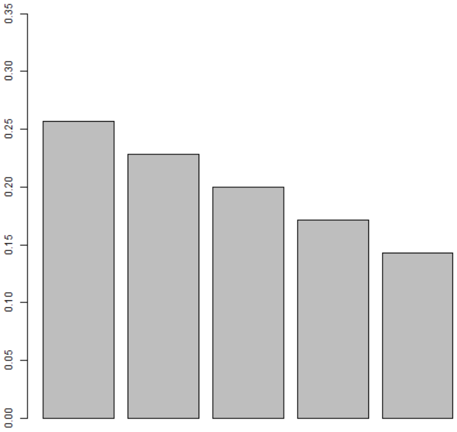

Figure 3.4: Underlying Distribution $\mathbf{q}$

And let $\alpha = 0.05$, number of iterations (for each sample size) $m = 100,000$, and sample sizes vary from 5 to $10,000$. The simulation results are summarized into Table

3.2 and Figure 3.5.

Table 3.2: Rejection Rates under H_a

Sample Size	5	10	50	10^2	10^3	10^4
T_l	0.0057	0.0310	0.2205	0.5531	1.0000	1.0000
T_p	0.9812	0.9416	0.9811	0.9956	1.0000	1.0000
T_z	1.0000	1.0000	0.9968	1.0000	1.0000	1.0000

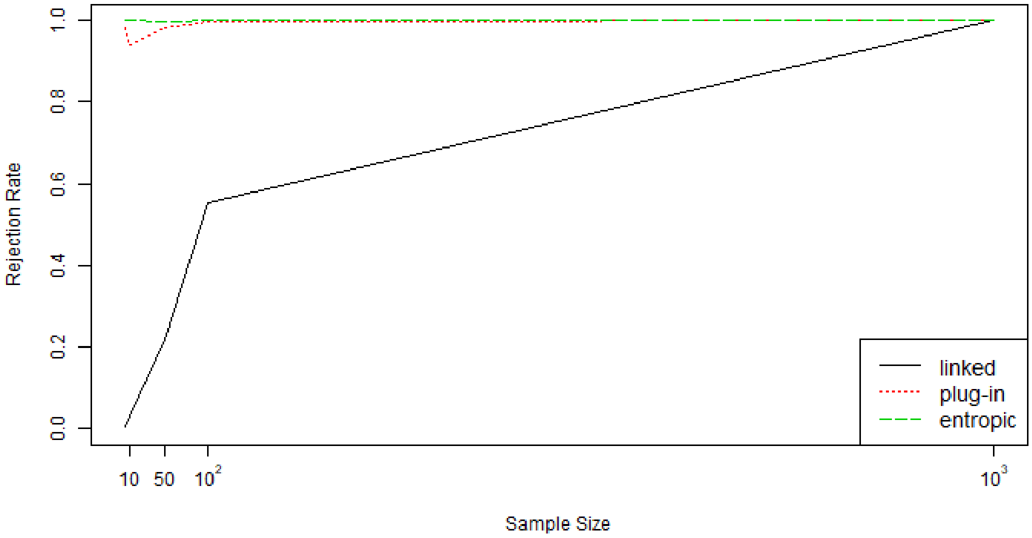

Figure 3.5: Rejection Rates under H_a

One can see the entropic test T_z and the plug-in test T_p behave closely, and both significantly outperform the linked Pearson's test T_l over all sample sizes, especially on a small sample ($n \leq 100$). As we mentioned above, due to the lack of test size control, those powers may be falsely high and can't be compared fairly. But this result still indicates some possible merits in the entropic test T_z and the plug-in test T_p under small samples.

3.2 A Heuristic Test

3.2.1 Testing Procedures

During the simulations in Section 3.1.2, we found two issues for the new Chi-squared tests. First, the minimum sample size to reach expected test size is too huge, *ie.*, the sampling distribution of test statistics is far away from Chi-squared distribution when sample size is small. The second issue is more about computation, but is quite realistic, that is we may easily encounter singularity errors when inverse matrices mentioned in (3.2) and (3.3), as cardinality increases ($K \geq 15$).

The first problem is well-known for many approximate tests [14], not just for our Chi-squared goodness-of-fit tests. The fundamental issue comes from the approximation to asymptotic distribution, which is derived by making the sample size big enough, and hence is unable to describe small sample phenomenon. So we may use exact test in substitution of approximate test, *ie.*, we use exact sampling distribution of test statistic to select critical values, instead of using asymptotic Chi-squared distribution. The exact distribution can be obtained by explicit formulation (for some simple cases, but very rare), or by large scale simulations [15] (in this dissertation we do $N = 100,000$ iterations for each simulation). The second problem doesn't challenge theoretical correctness of our method, but is fatal in real practice. After reviewed many literatures, we realized the inversion of large sparse matrix is still one of the biggest problems in computational algebra, so we consider modifying the test statistic instead.

Recall the expression of entropic test statistic in (3.3):

$$T_z = n[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)]^\tau (CA^\tau \Sigma AC^\tau)^{-1}[C(\mathbf{Z}^* - \boldsymbol{\eta}^*)] \tag{3.6}$$

This is in fact a weighted sum of squared differences between each pair of η_u and Z_v, also a measure of the distance between all η_us and all Z_vs. Generally speaking, goodness-of-fit measures the distance between underlying distribution and sampling distribution. Now from the entropic perspective, probability distributions can be characterized by entropic moments, so it's quite straightforward to measure that distance by the difference between η_us and Z_vs. Fortunately we have many mathematical instruments that can measure the distance between two functionals. Inspired the naive Euclidean distance, we construct the following test statistic:

$$T_h = (\mathbf{Z}^* - \boldsymbol{\eta}^*)^\tau (\mathbf{Z}^* - \boldsymbol{\eta}^*) = \sum_{u=2}^{K} (Z_u - \eta_u)^2 \tag{3.7}$$

And hence a new heuristic goodness-of-fit test, which rejects H_0 if

$$T_h = \sum_{u=2}^{K} (Z_u - \eta_u)^2 > C_\alpha(\mathbf{q}, n) \tag{3.8}$$

where for some $\alpha \in (0, 1)$, $C_\alpha(\mathbf{q}, n)$ is the $100(1 - \alpha)$ th percentile of statistic T_h's exact distribution under the null hypothesis with sample size n. $C_\alpha(\mathbf{q}, n)$ can be estimated from a large scale simulation.

This test is referred to in the subsequent text as the *heuristic test* T_h. The following corollary of Theorem 1 provides a theoretical support for this heuristic test.

Corollary 3. *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two probability distributions on the same finite alphabet* $\mathscr{X}$*. Then* $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ *if and only if*

$$\sum_{u=1}^{K} \left[\eta_u(\mathbf{p}) - \eta_u(\mathbf{q})\right]^2 = 0 \tag{3.9}$$

In fact, statistic T_h in (3.8) can also be viewed as a special case of statistic T_z in (3.6), if we force the weight matrix Σ^* to be the identity. Given the asymptotic multivariate normality of Z_vs (2.4), T_h in (3.8) is a linear combination of squared dependent normal variables. Unfortunately, a closed analytic expression for general sum of correlated Chi-squared variables is not yet known, which may be approximated efficiently using characteristic functions [16]. This is another reason why we choose to use exact test.

3.2.2 Simulations

We expect this heuristic test to be computationally more robust than the entropic test T_z, since its statistic has been greatly simplified, and don't lose too much power, as it also uses entropic moments. To evaluate the performance of T_h, we again run simulations.

In order to make the comparisons between different tests fair enough, we use exact test method, *ie.*, use simulated critical value for all of them. Here we have 4 different tests in total, linked Pearson's exact test ${T_l}^*$, plug-in exact test ${T_p}^*$, entropic exact test ${T_z}^*$ and the heuristic test T_h.

A great benefit of exact tests over approximate tests is that, the size can be easily controlled at any given level, since the critical values come from an exact distribution, so there is no need to examine test size anymore, we evaluate test power directly instead.

The simulations in this section are organized into 3 parts, the first of which follows similar studies as those in Section 3.1.2, comparing all 4 tests on small cardinality distributions ($K = 5$), and the second and third parts compare ${T_l}^*$ and T_h on large cardinality distributions ($K = 30$).

Simulation 1. Use the same settings as those in Section 3.1.2, sampling distribution $\mathbf{p}$ and underlying distribution $\mathbf{q}$ are given as follows,

$$\mathbf{p} = \left\{ \frac{5}{15}, \frac{4}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15} \right\} \tag{3.10}$$

$$\mathbf{q} = \left\{ \frac{9}{35}, \frac{8}{35}, \frac{7}{35}, \frac{6}{35}, \frac{5}{35} \right\} \tag{3.11}$$

And let $\alpha = 0.05$, number of iterations (for each sample size) $m = 100,000$, and sample sizes vary from 5 to $10,000$. Additionally, let $N = 100,000$ be the number of simulations to get $C_\alpha(\mathbf{q}, n)$ for each test. The simulation results are summarized into Table 3.3 and Figure 3.6.

Table 3.3: Rejection Rates under H_a

Sample Size	5	10	50	10^2	10^3	10^4
$T_l{}^*$	0.0731	0.0967	0.5517	0.8588	1.0000	1.0000
$T_p{}^*$	0.0730	0.1165	0.5911	1.0000	1.0000	1.0000
$T_z{}^*$	0.0732	0.0868	0.4185	0.7593	1.0000	1.0000
T_h	0.0739	0.1126	0.5131	0.8359	1.0000	1.0000

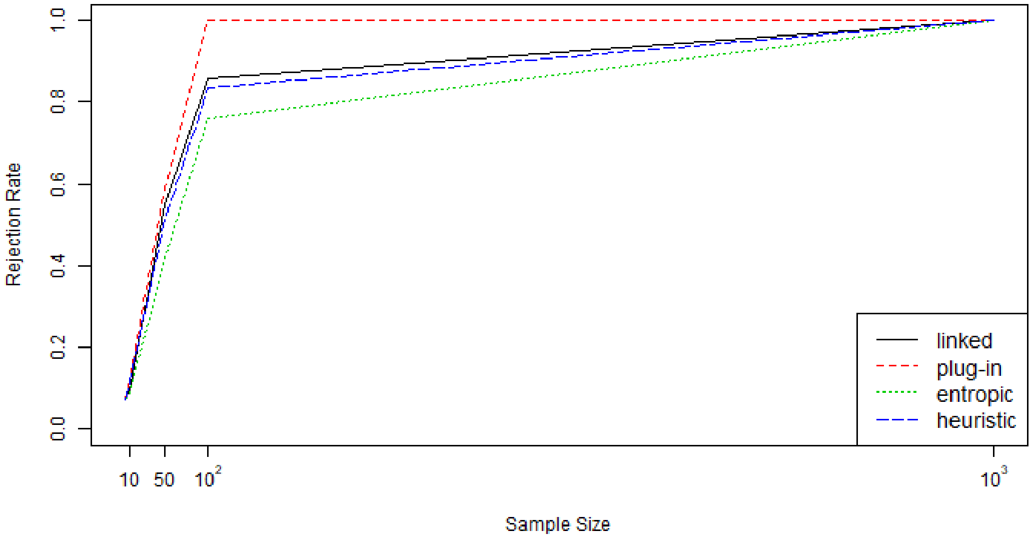

Figure 3.6: Rejection Rates under H_a

This result shows that all 4 tests perform closely in power sense, as the rejection rates reach 1 fast and smoothly when sample size exceeds $1,000$. The plug-in exact test is the best, as the linked Pearson's test and the heuristic test are almost the same. And that's a practical support for that on small cardinality distributions ($K = 5$), the heuristic test is totally qualified and comparable to those Chi-squared tests.

Besides, due to the computation issue as we stated above, the plug-in test and entropic test hardly work on large cardinality distributions. So in the second and third parts of this simulation, we only use the linked Pearson's test and the heuristic test for $K = 30$ cases.

Simulation 2. We pick the sampling probability distribution $\mathbf{p}$ and the underlying probability distribution $\mathbf{q}$ to be:

$$\mathbf{p} = \left\{ \frac{1}{30}, \frac{1}{30}, \cdots, \frac{1}{30} \right\} \tag{3.12}$$

$$\mathbf{q} = \left\{ \frac{1}{60}, \cdots, \frac{1}{60}, \frac{2}{60}, \cdots, \frac{2}{60}, \frac{3}{60}, \cdots, \frac{3}{60} \right\} \tag{3.13}$$

as shown in in Figure 3.7 and Figure 3.8.

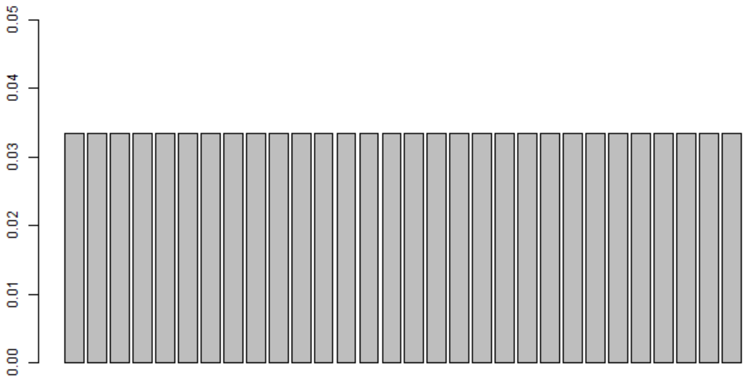

Figure 3.7: Sampling Distribution $\mathbf{p}$

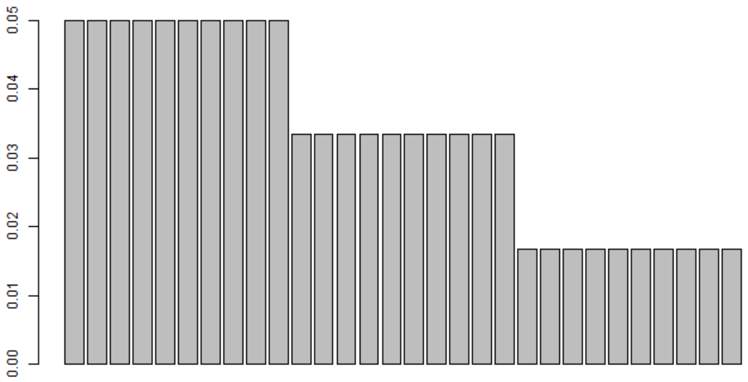

Figure 3.8: Underlying Distribution $\mathbf{q}$

And let $\alpha = 0.05$, number of iterations (for each sample size) $m = 100,000$, and sample sizes vary from 30 to 600. Additionally, let $N = 100,000$ be the number of simulations to get $C_\alpha(\mathbf{q}, n)$ for each test. The simulation results are summarized into Table 3.4 and Figure 3.9.

Table 3.4: Rejection Rates under H_a

Sample Size	30	60	90	150	300	400	500	600
T_l^*	0.0117	0.0026	0.0007	0.0001	0.0201	0.3561	0.8522	0.9862
T_h	0.0273	0.0874	0.2186	0.6010	0.9889	1.0000	1.0000	1.0000

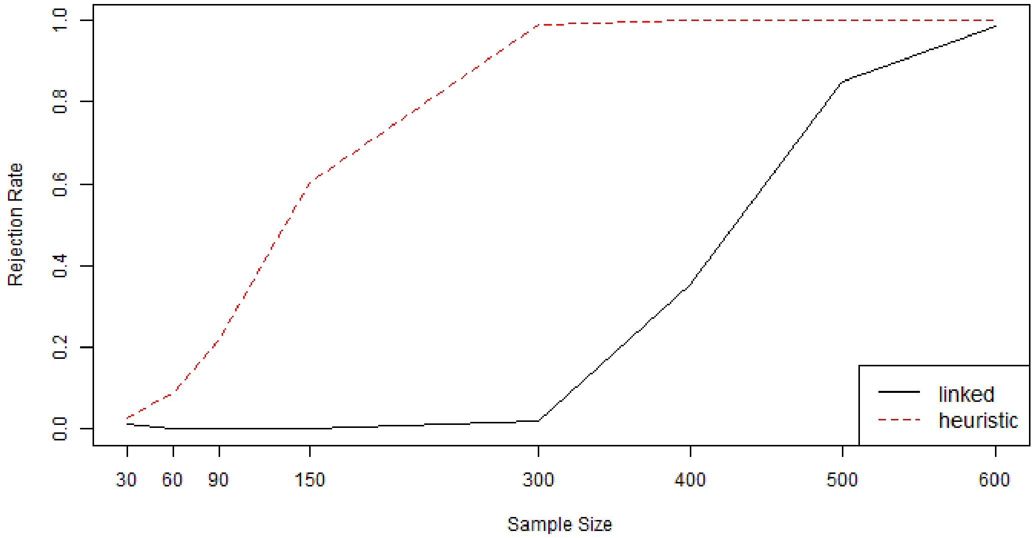

Figure 3.9: Rejection Rates under H_a

In this part, we give a real case where T_h thoroughly beats $T_l{}^*$ over all sample sizes. We take this example as an evidence of the merits of T_h, as well as of entropic perspective on alphabets.

Strictly speaking, we haven't proven any optimality of one test over others yet. But for all the simulations we've done so far (not just those presented here or in Appendix, but much more that we've done during the research), T_h never does significantly worse than $T_l{}^*$. In fact, for many cases, T_h is more likely to win, such as sample size is relatively small, or the underlying distribution has a thinner tail than the real sampling distribution. More interesting, as will be shown in next part of simulations, when the underlying distribution has a thicker tail than the sampling distribution instead, T_h doesn't lose power as compared to $T_l{}^*$, and we take this phenomena as another evidence of the merits of T_h and entropic perspective.

Simulation 3. We simply swap the values of $\mathbf{p}$ and $\mathbf{q}$, then redo Simulation 2.

$$\mathbf{p} = \left\{ \frac{1}{60}, \cdots, \frac{1}{60}, \frac{2}{60}, \cdots, \frac{2}{60}, \frac{3}{60}, \cdots, \frac{3}{60} \right\} \tag{3.14}$$

$$\mathbf{q} = \left\{ \frac{1}{30}, \frac{1}{30}, \cdots, \frac{1}{30} \right\} \tag{3.15}$$

The results are summarized into Table 3.5 and Figure 3.10.

Table 3.5: Rejection Rates under H_a

Sample Size	30	60	90	150	300
T_l^*	0.1465	0.3222	0.4955	0.8211	0.9972
T_h	0.1386	0.2885	0.4720	0.7946	0.9963

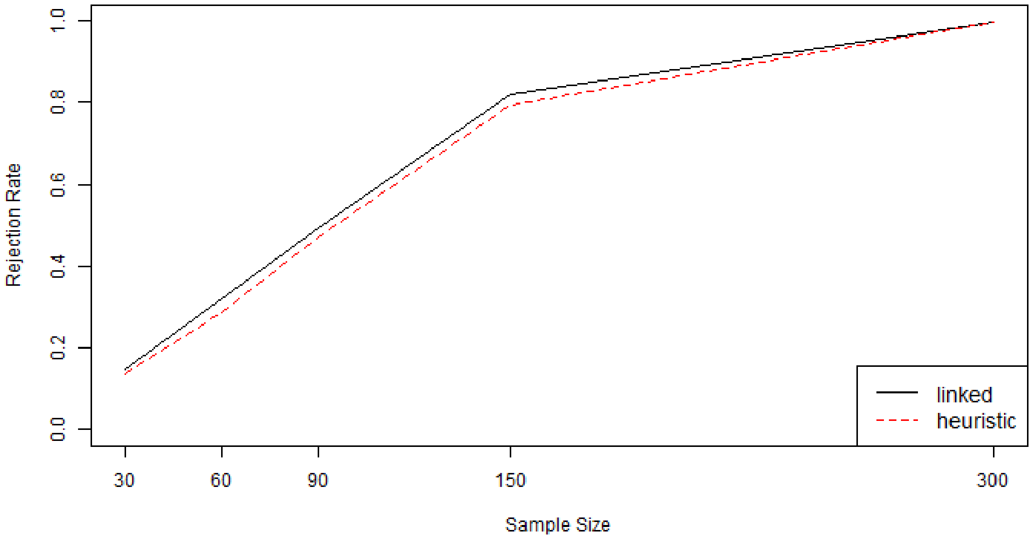

Figure 3.10: Rejection Rates under H_a

To summarize this chapter, we construct 7 goodness-of-fit tests in total, 3 approximate Chi-squared tests and 4 exact tests. Of course Chi-squared test are more computational efficient under large samples, but they may not work well for small or sparse samples, that's why we choose exact test method later. Unfortunately the plug-in test and entropic test face more tough issues like the matrix singularity problem, which cannot be easily solved yet. On large cardinality probability distributions, the simulation results show a great power advantage of the heuristic test over linked Pearson's exact test. One can find more simulation results in Appendix B.

We also believe the advantages of this heuristic test are mainly due to entropic perspective, *ie.*, the using of entropic moments and their estimator Z_vs. In this regard, one may define many similar statistics by imposing different metric or measures on entropic moments, which is completely up to the researcher's underlying interest and choice.

CHAPTER 4: Language Detection Example

4.1 Frequency Analysis

In cryptanalysis, frequency analysis is the study of the frequency of letters or groups of letters in a ciphertext. The method is used as an aid to breaking classical ciphers. Frequency analysis is based on the fact that, in any given stretch of written language, certain letters occur with varying frequencies. Moreover, there is a characteristic distribution of letters that is roughly the same for almost all samples of that language. For instance, given a section of English language, E, T, A and O are the most common, while Z, Q and X are rare [17]. In some ciphers, such properties of the natural language plaintext are preserved in the ciphertext, and these patterns have the potential to be exploited in a ciphertext-only attack.

But to make such decryption methods successful, the cryptanalyst must know a specific language in which the plaintext was written. Now the question is, given a piece of encrypted ciphertext, how to detect the language of its plaintext. Let's assume the encryption is done by some simple ciphers, *ie.*, the same letters in plaintext are still the same in ciphertext, and different letters in plaintext are still different in ciphertext. Given the fact that the distribution of letter frequencies vary cross different languages (Figure 4.1), we can use the letter frequencies counting from the ciphertext to detect the language of plaintext. This is indeed a goodness-of-fit test between the selected message and all possible languages, and it can also be viewed as a text language classifier.

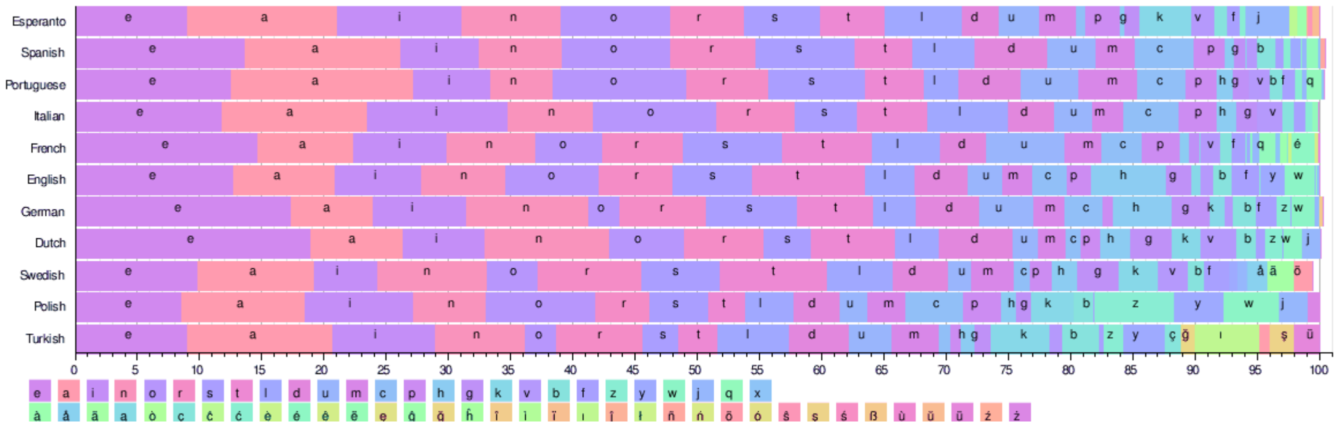

Figure 4.1: Frequency Distributions of 26 Most Common Latin Letters

Here we provide two examples to illustrate how this language detection method works. In each example, we select a piece of text from corpus, encrypt it by a Caesar cipher, then clean the ciphertext and break it down to bag of letters, count the letter frequencies, do the goodness-of-fit test between letter frequencies from the sample and that in all possible languages, and finally choose the one with largest p-value as possible source language.

The relative letter frequencies in 15 Latin languages were retrieved from Wikipedia [18] as references, and that version of data we've been using is included in Appendix C.

4.2 Testing Results

Example 1. We select the text of Martin Luther King's famous speech *"I Have a Dream"* as plaintext, with 885 words and 4781 letters in total. The full text was retrieved from BBS website [19]. The testing results are summarized into Table 4.1 and Table 4.2.

Table 4.1: Language Detection Statistic Values

English	**French**	**German**	**Spanish**	**Portuguese**
8.374921e-07	1.082819e-05	2.582115e-05	2.151521e-06	5.115412e-05
Esperanto	**Italian**	**Turkish**	**Swedish**	**Polish**
2.930989e-06	5.026936e-05	3.467886e-05	3.549134e-05	1.980893e-04
Dutch	**Danish**	**Icelandic**	**Finnish**	**Czech**
1.839462e-04	2.245837e-06	1.656464e-04	2.403325e-05	1.255422e-04

Table 4.2: Language Detection p-values

English	**French**	**German**	**Spanish**	**Portuguese**
0.4016	0.0102	0.0017	0.2106	0.0000
Esperanto	**Italian**	**Turkish**	**Swedish**	**Polish**
0.1130	0.0000	0.0000	0.0000	0.0000
Dutch	**Danish**	**Icelandic**	**Finnish**	**Czech**
0.0000	0.2996	0.0000	0.0000	0.0000

Our test shows English is the most probable language among all 15 candidates, and it does discover the truth. Someone may notice that, the observed significance levels of Spanish and Danish are also high. It might be due to the timeliness of letter frequencies data. Because the relative letter frequencies data we retrieved from Wikipedia were collected in year 2014, and this famous speech was given in year 1963, so 50 years may make a big difference between "old" English and current English. The second example can be a side evidence of this guess.

Example 2. We select the text of Donald Trump's inauguration speech as plaintext, with 1427 words and 8077 letters in total. The full text was retrieved from CNN website [20]. Compared with the first text sample, this speech was given in year 2017, and since it is more "modern", we expect to see a more significant result from the test. The results are summarized into Table 4.3 and Table 4.4.

Table 4.3: Language Detection Statistic Values

English	**French**	**German**	**Spanish**	**Portuguese**
7.536048e-08	1.992632e-05	3.894047e-05	6.967875e-06	6.929651e-05
Esperanto	**Italian**	**Turkish**	**Swedish**	**Polish**
8.342205e-06	6.833214e-05	2.225812e-05	2.286667e-05	1.664068e-04
Dutch	**Danish**	**Icelandic**	**Finnish**	**Czech**
2.169609e-04	6.713070e-06	1.367873e-04	3.694338e-05	1.005868e-04

Table 4.4: Language Detection p-values

English	**French**	**German**	**Spanish**	**Portuguese**
0.7493	0.0001	0.0000	0.0040	0.0000
Esperanto	**Italian**	**Turkish**	**Swedish**	**Polish**
0.0005	0.0000	0.0000	0.0000	0.0000
Dutch	**Danish**	**Icelandic**	**Finnish**	**Czech**
0.0000	0.0179	0.0000	0.0000	0.0000

Not surprisingly, English, as the truth, is again detected by the heuristic goodness-of-fit test. And English is the only one producing a high p-value, with all other p-values are almost zero. We take this example as a strong support for practical utility of the heuristic test, and entropic perspective.

CHAPTER 5: An R Package

During the research on entropic perspective, we find all computations related with entropic moments $\zeta_{u,v}$s and their estimators $Z_{u,v}$s are quite time consuming, as they involve a lot of huge combinatorics. To improve the computation efficiency and consequently to save more time for thinking instead of coding, we build an R package named as "Entropic". All computation intensive functions are written in C++, and some auxiliary functions are written in R. The source code of several key functions can be downloaded at [https://webpages.uncc.edu/cchen55/entropic/Entropic.zip].

Here is a brief introduction of some core functions:

- tf1(sample) returns the Truing's Formula for a given sample;
- entropy(prob, k) returns the entropy of a given distribution of length k;
- zeta1(prob, k, v) returns the $\zeta_{1,v}$ value for a given v;
- zeta1f(prob, k, vm) returns a vector of all $\zeta_{1,v}$ values for $v \leq vm$;
- eta(prob, k, u) returns the η_u value for a given u;
- etaf(prob, k, um) returns a vector of all η_u values for $u \leq um$;
- z1(obs, k, n, v) returns the $Z_{1,v}$ value for a given v;
- z1f(obs, k, n) returns a vector of all $Z_{1,v}$ values for $v < n$;
- entropyz(obs, k, n) returns the entropic estimator of entropy $\hat{H}_z$.

CHAPTER 6: Conclusion

In modern data science, many challenges arise from high dimensionality and data complexity. To handle those problems in a broader perspective, one may view them as on a countable discrete sample space, non-metricized, non-ordinal, not completely prescribed (*ie., alphabet*), but with distinguishable elements (*ie., letters*). Entropic basis works well as a new characterization of probability distributions, while many traditional statistical concepts like cumulative distribution function and moments may not even exist.

Entropic perspective, in the forms of entropic basis, builds a bridge between seen and unseen, and provides a whole new instrument to interpret non-ordinal and sparse data. One can see the power of this new perspective in many applications, such as estimation of entropy and mutual information [21] [22] [23] [24], estimation of diversity indices [11] [10] [25], independence test [26], etc.

This dissertation focuses on goodness-of-fit test under permutations on alphabets. Equipped with entropic perspective, we propose 3 approximate tests and 4 exact tests. For approximate tests, asymptotic Chi-squared distributions are derived. For exact tests, computation efficiency and size control are greatly improved. At the end, the real data example, language detection classifier demonstrates a great potential of this methodology in practice.

REFERENCES

[1] Z. Zhang, *Statistical Implications of Turing's Formula.* Wiley, 2016.

[2] R. Bellman, R. Corporation, and K. M. R. Collection, *Dynamic Programming.* Rand Corporation research study, Princeton University Press, 1957.

[3] T. R. Read and N. A. Cressie, *Goodness-of-fit statistics for discrete multivariate data.* Springer Science & Business Media, 2012.

[4] A. Kolmogorov, "Sulla determinazione empirica di una lgge di distribuzione," *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[5] T. W. Anderson, D. A. Darling, *et al.*, "Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes," *The annals of mathematical statistics*, vol. 23, no. 2, pp. 193–212, 1952.

[6] K. Pearson, "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," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[7] J. H. McDonald, *Handbook of biological statistics*, vol. 2. 2009.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[9] T. Cover and J. Thomas, *Elements of Information Theory.* Wiley, 2012.

[10] Z. Zhang and M. Grabchak, "Entropic representation and estimation of diversity indices," *Journal of Nonparametric Statistics*, vol. 28, no. 3, pp. 563–575, 2016.

[11] Z. Zhang and J. Zhou, "Re-parameterization of multinomial distributions and diversity indices," *Journal of Statistical Planning and Inference*, vol. 140, no. 7, pp. 1731–1738, 2010.

[12] Z. Zhang *et al.*, "Domains of attraction on countable alphabets," *Bernoulli*, vol. 24, no. 2, pp. 873–894, 2018.

[13] S. Molchanov, Z. Zhang, and L. Zheng, "Entropic moments and domains of attraction on countable alphabets," *Mathematical Methods of Statistics*, vol. 27, no. 1, pp. 60–70, 2018.

[14] R. A. Fisher, *Statistical methods for research workers.* Genesis Publishing Pvt Ltd, 2006.

[15] B. Dimitrov, D. Green, V. Rykov, and P. Stanchev, "On statistical hypothesis testing via simulation method," 01 2003.

[16] J. Bausch, “On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua,” *Journal of Physics A: Mathematical and Theoretical*, vol. 46, no. 50, p. 505202, 2013.

[17] S. Singh, “The black chamber: Hints and tips,” 2010. [Online; Retrieved October 26, 2010].

[18] Wikipedia Contributors, “Letter frequency,” 2019. [Online; Accessed March 13, 2019].

[19] J. Martin Luther King, “I have a dream,” 1963. [BBC; Accessed March 13, 2019].

[20] D. Trump, “Inaugural address: Trump’s full speech,” 2017. [CNN; Accessed March 13, 2019].

[21] C. Chen, M. Grabchak, A. Stewart, J. Zhang, and Z. Zhang, “Normal laws for two entropy estimators on infinite alphabets,” *Entropy*, vol. 20, no. 5, p. 371, 2018.

[22] Z. Zhang, “Entropy estimation in turing’s perspective,” *Neural computation*, vol. 24, no. 5, pp. 1368–1389, 2012.

[23] Z. Zhang, “Asymptotic normality of an entropy estimator with exponentially decaying bias,” *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 504–508, 2012.

[24] Z. Zhang and A. M. Stewart, “Estimation of standardized mutual information,” tech. rep., UNC Charlotte Technical Report, 2016.

[25] Z. Zhang, C. Chen, and J. Zhang, “Estimation of population size in entropic perspective,” *Communications in Statistics - Theory and Methods*, 2018.

[26] J. Zhang and C. Chen, “On "a mutual information estimator with exponentially decaying bias" by zhang and zheng,” *Statistical applications in genetics and molecular biology*, vol. 17, 03 2018.

[27] C. B. Garcia and T. Y. Li, “On the number of solutions to polynomial systems of equations,” *SIAM Journal on Numerical Analysis*, vol. 17, no. 4, pp. 540–546, 1980.

[28] J. L. Doob, “The limiting distributions of certain statistics,” *The Annals of Mathematical Statistics*, vol. 6, no. 3, pp. 160–169, 1935.

APPENDIX A: Proofs

A.1 Proofs in Section 2.1

To prove Theorem 1, a result by Garcia and Li (1980) [27] is stated as Lemma 4 below.

A.1.1 Lemma 4

Toward stating the lemma, let $\mathbf{z} = \{z_1, \cdots, z_n\}$ be a multivariate variable in the n-dimensional complex space $\mathbb{C}^n$. Consider a system of n polynomial equations where each equation is a summation with terms of the following form

$$az_1^{r_1} z_2^{r_2} \cdots z_n^{r_n} \tag{A.1}$$

equal to zero, where the sum of non-negative integers $r_1 + \cdots + r_n$ is the degree of the additive term, and $a \in \mathbb{C}$ is the coefficient of the term. For the i^{th} equation, let q_i denote the highest degree among all the additive terms. Let the term coefficients of all equations be denoted as $\mathbf{a} = \{a_{i,j}; i = 1, \cdots, n \text{ and } j = 1, \cdots, m\}$ for some positive integer m which depends on $\max\{q_i; i = 1, \cdots, n\}$. For notation simplicity, let $\mathbf{a}_i = \{a_{i,j}; j = 1, \cdots, m\}$. Let $\mathbf{P}(\mathbf{z}, \mathbf{a}) = 0$ denote the polynomial equation system. The i^{th} equation in $\mathbf{P}(\mathbf{z}, \mathbf{a}) = 0$ has the form of $P_i(\mathbf{z}, \mathbf{a}_i) = 0$ with the left hand side being a polynomial of degree q_i. For each i, removing all the additive terms of degrees less than q_i in $P_i(\mathbf{z}, \mathbf{a}_i)$ results in a homogeneous polynomial $Q_i(\mathbf{z}, \mathbf{a}_i)$ of degree q_i, and setting it to zero gives an adjusted equation, $Q_i(\mathbf{z}, \mathbf{a}_i) = 0$. Let the adjusted system be denoted as $\mathbf{Q}(\mathbf{z}, \mathbf{a}) = 0$.

Lemma 4. *Let $\mathbf{P}(\mathbf{z}, \mathbf{a}) = 0$ be given and let $\mathbf{Q}(\mathbf{z}, \mathbf{a}) = 0$ be its corresponding highest order system of equations. $\mathbf{P}(\mathbf{z}, \mathbf{a}) = 0$ has exactly $q = \prod_{i=1}^{n} q_i$ solutions (counting multiplicity) if $\mathbf{Q}(\mathbf{z}, \mathbf{a}) = 0$ has only the trivial solution $\mathbf{z} = \{0, \cdots 0\}$.*

A.1.2 Proof of Theorem 1

Proof. Given $\mathbf{p} = \{p_k; k = 1, \cdots, K\}$, $\boldsymbol{\eta} = \{\eta_u; u = 1, \cdots, K\}$ is uniquely determined. It suffices to show that $\boldsymbol{\eta}$ uniquely determines $\mathbf{p}$. Toward that end, consider the system of equations in (A.2) and its adjusted system in (A.3), denoted respectively as $\mathbf{P}(\mathbf{z}, \mathbf{a}) = 0$ and $\mathbf{Q}(\mathbf{z}, \mathbf{a}) = 0$.

$$\begin{cases} \sum_{k=1}^{K} p_k & = & \eta_1 \\ \sum_{k=1}^{K} p_k^2 & = & \eta_2 \\ & \vdots & \\ \sum_{k=1}^{K} p_k^K & = & \eta_K \end{cases} \tag{A.2}$$

$$\begin{cases} \sum_{k=1}^{K} p_k & = & 0 \\ \sum_{k=1}^{K} p_k^2 & = & 0 \\ & \vdots & \\ \sum_{k=1}^{K} p_k^K & = & 0 \end{cases} \tag{A.3}$$

Clearly $\mathbf{p} = \{p_1, \cdots, p_K\}$ is a solution to (A.2) and so is every permutated $\mathbf{p}$. Therefore counting multiplicity, there are at least $q = \prod_{i=1}^{K} q_i = K!$ solutions to (A.2) and all these solutions share the same $\mathbf{p}_\downarrow$. It only remains to show that there are no other solutions. By Lemma 4, it is desired to show that the system (A.3) only has trivial solution of $p_k = 0$ for every $k = 1, \cdots, K$.

Toward that end, consider the linear system in $u_1, u_2, \cdots, u_n$:

$$\begin{cases} 1 \cdot u_1 + 1 \cdot u_2 + \cdots + 1 \cdot u_n & = & 0 \\ x_1 \cdot u_1 + x_2 \cdot u_2 + \cdots + x_n \cdot u_n & = & 0 \\ & \vdots & \\ x_1^{n-1} \cdot u_1 + x_2^{n-1} \cdot u_2 + \ldots + x_n^{n-1} \cdot u_n & = & 0 \end{cases} \quad \text{or} \quad A \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \emptyset$$

where

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ & \cdots & & \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{pmatrix}$$

Assuming not all x_i are 0, that is, the system has the non-trivial solution $u_1 = x_1, u_2 = x_2, \cdots, u_n = x_n$, so its determinant, $\det(A)$, must be 0. But the respective determinant is a Vandermonde determinant which evaluates to $\prod_{1\le i<j\le n}(x_j - x_i)$, so it can only be zero if $x_i = x_j$ for some pair $i \neq j$.

Assume without loss of generality that $x_{n-1} = x_n$, then consider:

$$\begin{cases} 1\cdot u_1 + 1\cdot u_2 + \cdots + 1\cdot u_{n-1} &= 0 \\ x_1\cdot u_1 + x_2\cdot u_2 + \cdots + x_{n-1}\cdot u_{n-1} &= 0 \\ & \vdots \\ x_1^{n-2}\cdot u_1 + x_2^{n-2}\cdot u_2 + \cdots + x_{n-1}^{n-2}\cdot u_{n-1} &= 0 \end{cases}$$

Again, if not all x_i are 0, then $u_1 = x_1, u_2 = x_2, \cdots, u_{n-2} = x_{n-2}, u_{n-1} = 2x_{n-1}$ is a non-trivial solution, which implies that another pair of $x_i = x_j$, $i \neq j$.

It follows by induction that all x_i must be equal, and therefore all must be 0. □

A.2 Proofs in Section 2.2

A.2.1 Proof of Lemma 2

Proof. Let $\mathbf{p}_- = (p_1, \cdots, p_{K-1})^\tau$ and $\hat{\mathbf{p}}_- = (\hat{p}_1, \cdots, \hat{p}_{K-1})^\tau$. Noting the fact that $\sqrt{n}(\hat{\mathbf{p}}_- - \mathbf{p}_-) \xrightarrow{L} N(\mathbf{0}, \Sigma)$, (2.5) is verified by a straight forward application of the multivariate delta method [28]. Since Σ is a matrix of full rank, it only remains to show that A has rank r.

Toward that end, consider first the case that all p_ks are distinct, that is, $r = K-1$. Suppose there exists $(w_1, \cdots, w_{K-1})$ such that, $w_1(p_1^i - p_K^i) + w_2(p_2^i - p_K^i) + \cdots +$

$w_{K-1}(p^i_{K-1} - p^i_K) = 0$ for every i, $i = 1, \cdots, K-1$, that is,

$$\begin{cases} w_1 p_1^1 + w_2 p_2^1 + \cdots + w_{K-1} p_{K-1}^1 - (\sum_{j=1}^{K-1} w_j) p_K^1 & = \quad 0 \\ w_1 p_1^2 + w_2 p_2^2 + \cdots + w_{K-1} p_{K-1}^2 - (\sum_{j=1}^{K-1} w_j) p_K^2 & = \quad 0 \\ & \cdots \\ w_1 p_1^{K-1} + w_2 p_2^{K-1} + \cdots + w_{K-1} p_{K-1}^{K-1} - (\sum_{j=1}^{K-1} w_j) p_K^{K-1} & = \quad 0 \end{cases} \tag{A.4}$$

Letting $w_K = -\sum_{j=1}^{K-1} w_j$, $\sum_{j=1}^{K} w_j = 0$ by definition and this equation can be added to the system in (A.4) to obtain an equivalent system in $w_1, \cdots, w_K$ below.

$$\begin{cases} 1 \cdot w_1 + 1 \cdot w_2 + \cdots + 1 \cdot w_{K-1} + 1 \cdot w_K & = \quad 0 \\ p_1^1 \cdot w_1 + p_2^1 \cdot w_2 + \cdots + p_{K-1}^1 \cdot w_{K-1} + p_K^1 \cdot w_K & = \quad 0 \\ & \cdots \\ p_1^{K-1} \cdot w_1 + p_2^{K-1} \cdot w_2 + \cdots + p_{K-1}^{K-1} \cdot w_{K-1} + p_K^{K-1} \cdot w_K & = \quad 0 \end{cases} \tag{A.5}$$

If $(w_1, \cdots, w_{K-1}) \neq \mathbf{0}$ then $(w_1, \cdots, w_K) \neq \mathbf{0}$, which implies that the Vandermonde determinant associated with (A.5) must be zero, which evaluates to $\prod_{1 \leq i < j \leq K}(p_i - p_j)$, so it can only be zero if $p_i = p_j$ for some pair $i \neq j$, which contradicts the assumption. It follows that A is of full rank, $r = K-1$, if all p_ks are distinct.

Next consider the case there are $r+1$ distinct values in $\{p_1, \cdots, p_K\}$ where r is an integer such that $0 \leq r \leq K-1$. In this case, any set of more than r columns of A in (2.7) are linearly independent. This claim may be seen in two scenarios. First, if p_K has multiplicity 1, say $m_K = 1$, then the $K-1$ columns of A include exactly r distinct columns. Therefore any subset of more than r of these columns must contain at least a pair of identical columns. Second, if p_K has multiplicity greater than 1, that is, $m_K \geq 2$, then the $K-1$ columns of A include exactly $m_K - 1 \geq 1$ all-zero columns and other r non-zero distinct columns. In this case, any subset of more r columns either contains at least one pair of identical columns or an all-zero column.

That is to say that the rank of A is at most r. It suffices to show that the said rank is at least r. Toward that end, consider r distinct columns of A, (A.6). Without loss of generality, suppose these columns are for $j = 1, \cdots, r$.

$$A_1 = \begin{pmatrix} (p_1 - p_K) & (p_2 - p_K) & \cdots & (p_r - p_K) \\ (p_1^2 - p_K^2) & (p_2^2 - p_K^2) & \cdots & (p_r^2 - p_K^2) \\ \vdots & \vdots & \ddots & \vdots \\ (p_1^{K-1} - p_K^{K-1}) & (p_2^{K-1} - p_K^{K-1}) & \cdots & (p_r^{K-1} - p_K^{K-1}) \end{pmatrix} \tag{A.6}$$

The desired independence of the columns of (A.6) is established by showing that the columns of (A.7) are linearly independent. Consider a $r \times r$ sub-matrix of (A.6) below.

$$A_2 = \begin{pmatrix} (p_1 - p_K) & (p_2 - p_K) & \cdots & (p_r - p_K) \\ (p_1^2 - p_K^2) & (p_2^2 - p_K^2) & \cdots & (p_r^2 - p_K^2) \\ \vdots & \vdots & \ddots & \vdots \\ (p_1^r - p_K^r) & (p_2^r - p_K^r) & \cdots & (p_r^r - p_K^r) \end{pmatrix}_{r \times r} \tag{A.7}$$

Suppose the columns of (A.7) are linearly dependent, then there exists $(w_1, \cdots, w_r) \neq \mathbf{0}$ such that

$$\begin{cases} w_1 p_1^1 + w_2 p_2^1 + \cdots + w_r p_r^1 - (\sum_{j=1}^{r-1} w_j) p_K^1 & = & 0 \\ w_1 p_1^2 + w_2 p_2^2 + \cdots + w_r p_r^2 - (\sum_{j=1}^{r-1} w_j) p_K^2 & = & 0 \\ & \cdots & \\ w_1 p_1^r + w_2 p_2^r + \cdots + w_r p_r^r - (\sum_{j=1} w_r) p_K^r & = & 0 \end{cases} \tag{A.8}$$

Letting $w_{r+1} = -\sum_{j=1}^{r-1} w_j$, $\sum_{j=1}^{r+1} w_j = 0$ by definition and this equation can be added to the system in (A.8) to obtain an equivalent system in $w_1, \cdots, w_K$ below.

$$\begin{cases} 1 \cdot w_1 + 1 \cdot w_2 + \cdots + 1 \cdot w_r + 1 \cdot w_{r+1} & = \quad 0 \\ p_1^1 \cdot w_1 + p_2^1 \cdot w_2 + \cdots + p_r^1 \cdot w_r + p_K^1 \cdot w_{r+1} & = \quad 0 \\ p_1^2 \cdot w_1 + p_2^2 \cdot w_2 + \cdots + p_r^2 \cdot w_r + p_K^2 \cdot w_{r+1} & = \quad 0 \\ & \cdots \\ p_1^{r-1} \cdot w_1 + p_2^{r-1} \cdot w_2 + \cdots + p_r^r \cdot w_r + p_K^r \cdot w_{r+1} & = \quad 0 \end{cases} \tag{A.9}$$

If the system in (A.9) has an not all-zero solution in $w_1, \cdots, w_{r+1}$, then its associated determinant must be zero, which is a Vandermonde determinant and evaluates to $\prod_{i,j=1,\cdots,r,K;i<j}(p_i - p_j)$, so it can only be zero if $p_i = p_j$ for some pair $i \neq j$, which contradicts the assumption. It follows that A is of rank, r, if $r+1$ of p_ks are distinct.

□

A.2.2 Proof of Lemma 3

Proof. Since

$$
\begin{aligned}
Z_{1,v} - \hat{\zeta}_v &= Z_{1,v} - \sum_{k=1}^{K} \hat{p}_k (1-\hat{p}_k)^v \\
&= \sum_{k=1}^{K} \hat{p}_k \prod_{j=1}^{v} \left(1 - \frac{Y_k - 1}{n-j}\right) - \sum_{k=1}^{K} \hat{p}_k (1-\hat{p}_k)^v \\
&= \sum_{k=1}^{K} \hat{p}_k \left[\prod_{j=1}^{v} \left(1 - \frac{Y_k - 1}{n-j}\right) - \prod_{j=1}^{v} (1-\hat{p}_k)\right] \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left[\frac{\prod_{j=1}^{v}\left(1 - \frac{Y_k-1}{n-j}\right)}{\prod_{j=1}^{v}\left(1-\hat{p}_k\right)} - 1\right]\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{\frac{\prod_{j=1}^{v}\left[1 - \frac{j-1}{n(1-\hat{p}_k)}\right]}{\prod_{j=1}^{v}\left(1-\frac{j}{n}\right)} - 1\right\}\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{\frac{\prod_{j=0}^{v-1}\left[1 - \frac{j}{n(1-\hat{p}_k)}\right]}{\prod_{j=0}^{v-1}\left(1-\frac{j+1}{n}\right)} - 1\right\}\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{\prod_{j=0}^{v-1}\left[\frac{1 - \frac{j}{n(1-\hat{p}_k)}}{1-\frac{j+1}{n}}\right] - 1\right\}\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{\prod_{j=0}^{v-1}\left[1 + \frac{j(1-\hat{p}_k) - (j-1)}{n(1-\hat{p}_k) - j(1-\hat{p}_k)}\right] - 1\right\}\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{1 + \sum_{j=0}^{v-1}\left[\frac{j(1-\hat{p}_k) - (j-1)}{n(1-\hat{p}_k) - j(1-\hat{p}_k)}\right] + \mathcal{O}_p(n^{-2}) - 1\right\}\right\} \\
&= \sum_{k=1}^{K} \left\{\hat{p}_k \left(1-\hat{p}_k\right)^v \left\{\sum_{j=0}^{v-1}\left[\frac{j(1-\hat{p}_k) - (j-1)}{n(1-\hat{p}_k) - j(1-\hat{p}_k)}\right] + \mathcal{O}_p(n^{-2})\right\}\right\}
\end{aligned}
$$

Therefore

$$\begin{aligned}
&n\left[Z_{1,v}-\sum_{k=1}^{K}\hat{p}_k(1-\hat{p}_k)^v\right]\\
&=\sum_{k=1}^{K}\left\{\hat{p}_k\,(1-\hat{p}_k)^v\left\{\sum_{j=0}^{v-1}\left[\frac{j(1-\hat{p}_k)-(j-1)}{(1-\hat{p}_k)-j(1-\hat{p}_k)/n}\right]+\mathcal{O}_p(n^{-1})\right\}\right\}\\
&\xrightarrow{p}\sum_{k=1}^{K}\left\{p_k\,(1-p_k)^v\left\{\sum_{j=0}^{v-1}\left[\frac{j(1-p_k)-(j-1)}{(1-p_k)}\right]\right\}\right\}\\
&=\sum_{k=1}^{K}\left\{p_k\,(1-p_k)^{v-1}\left\{\sum_{j=0}^{v-1}[j(1-p_k)-(j-1)]\right\}\right\}\\
&=\sum_{k=1}^{K}\left\{p_k\,(1-p_k)^{v-1}\left[v-\left(\sum_{j=0}^{v-1}j\right)p_k\right]\right\}\\
&=v\zeta_{1,v-1}-\frac{v(v-1)}{2}\zeta_{2,v-1} \qquad \text{(A.10)}
\end{aligned}$$

□

A.2.3 Proof of Theorem 2

Proof. Since $\sqrt{n}(\mathbf{Z}^*-\boldsymbol{\eta}^*)=\sqrt{n}(\mathbf{Z}^*-\hat{\boldsymbol{\eta}}^*)+\sqrt{n}(\hat{\boldsymbol{\eta}}^*-\boldsymbol{\eta}^*)$, it suffices to show that $\sqrt{n}(\mathbf{Z}^*-\hat{\boldsymbol{\eta}}^*)\xrightarrow{p}0$. Toward that end, noting the following two easily verifiable re-expressions of Z_u and $\hat{\eta}_u$,

$$Z_u=\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^iZ_{1,i}\qquad\text{and}\qquad\hat{\eta}_u=\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i\hat{\zeta}_{1,i}$$

then by Lemma 3,

$$\begin{aligned}
\sqrt{n}(Z_u-\hat{\eta}_u)&=\sqrt{n}\left[\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^iZ_{1,i}-\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i\hat{\zeta}_i\right]\\
&=\sum_{i=0}^{u-1}\binom{u-1}{i}(-1)^i\sqrt{n}(Z_{1,i}-\hat{\zeta}_i)\xrightarrow{p}0
\end{aligned}$$

□

APPENDIX B: More Simulation Results

We define a family of 10 different distributions, as described in (B.1) and shown in Figure B.1.

$$\mathbf{p_i} = (\underbrace{\frac{1-\epsilon_i}{30}, \cdots,}_{10} \underbrace{\frac{1}{30}, \cdots,}_{10} \underbrace{\frac{1+\epsilon_i}{30}, \cdots}_{10}), \qquad \epsilon_i = \frac{i-1}{10}, \qquad i = 1, 2, \cdots, 10 \qquad \text{(B.1)}$$

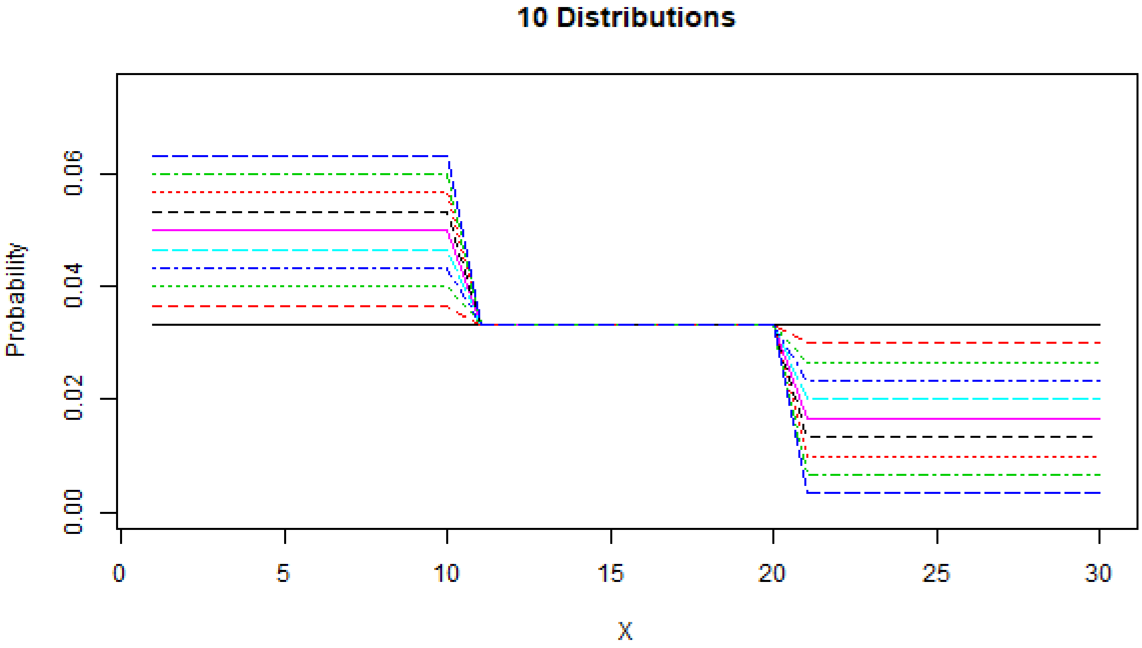

Figure B.1: A Family of 10 Probability Distributions

Now for all possible combinations of $\mathbf{p_i}$ and $\mathbf{p_j}$ where $i, j \in \{1, 2 \cdots, 10\}$, let $\mathbf{p_i} = \mathbf{p}$ be the sampling distribution, $\mathbf{p_j} = \mathbf{q}$ be the underlying distribution, $\alpha = 0.05$, number of iterations (for each sample size) $m = 5,000$, and sample sizes vary from 30 to 90. Additionally, let $N = 1,000,000$ be the number of simulations to get critical for each test. The test rejection rates are summarized into Tables B.1, B.2 and B.3.

Table B.1: Rejection Rates When $n = 30$

p|q	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10
i=1	0.0451	0.0470	0.0376	0.0343	0.0365	0.0360	0.0285	0.0655	0.1378	0.2892
i=2	0.0492	0.0440	0.0414	0.0333	0.0322	0.0355	0.0301	0.0670	0.1330	0.2763
i=3	0.0550	0.0551	0.0495	0.0406	0.0371	0.0340	0.0318	0.0563	0.1116	0.2611
i=4	0.0682	0.0697	0.0636	0.0447	0.0409	0.0344	0.0272	0.0489	0.0940	0.2291
i=5	0.0985	0.0897	0.0848	0.0657	0.0476	0.0393	0.0313	0.0373	0.0737	0.1830
i=6	0.1378	0.1189	0.1175	0.0926	0.0666	0.0496	0.0345	0.0383	0.0576	0.1414
i=7	0.2002	0.1681	0.1659	0.1379	0.0996	0.0704	0.0481	0.0337	0.0440	0.1016
i=8	0.2700	0.2367	0.2257	0.1883	0.1435	0.1087	0.0722	0.0509	0.0378	0.0638
i=9	0.3641	0.3316	0.3156	0.2670	0.2081	0.1581	0.1043	0.0714	0.0493	0.0506
i=10	0.4892	0.4406	0.4368	0.3810	0.3058	0.2362	0.1632	0.1178	0.0780	.05290

Table B.2: Rejection Rates When $n = 60$

p|q	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10
i=1	0.0518	0.0420	0.0401	0.0359	0.0453	0.0808	0.1968	0.3988	0.6662	0.8669
i=2	0.0543	0.0482	0.0395	0.0392	0.0417	0.0827	0.1933	0.3933	0.6422	0.8615
i=3	0.0707	0.0634	0.0473	0.0397	0.0399	0.0631	0.1484	0.3337	0.5883	0.8120
i=4	0.1149	0.0961	0.0706	0.0502	0.0360	0.0452	0.1086	0.2453	0.4980	0.7467
i=5	0.1759	0.1656	0.1200	0.0876	0.0491	0.0391	0.0661	0.1708	0.3681	0.6435
i=6	0.2848	0.2666	0.2021	0.1473	0.0883	0.0487	0.0473	0.0915	0.2366	0.4942
i=7	0.4460	0.4175	0.3511	0.2622	0.1653	0.0947	0.0508	0.0479	0.1243	0.3182
i=8	0.6458	0.6082	0.5326	0.4274	0.3044	0.1847	0.0928	0.0533	0.0611	0.1690
i=9	0.8264	0.7981	0.7261	0.6371	0.4985	0.3492	0.1931	0.0993	0.0489	0.0672
i=10	0.9504	0.9350	0.8986	0.8366	0.7278	0.5622	0.3823	0.2080	0.0989	0.0481

Table B.3: Rejection Rates When $n = 90$

p|q	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10
i=1	0.0524	0.0471	0.0376	0.0408	0.0980	0.2227	0.4776	0.7763	0.9403	0.9935
i=2	0.0560	0.0557	0.0385	0.0394	0.0848	0.2039	0.4507	0.7566	0.9273	0.9891
i=3	0.0806	0.0749	0.0508	0.0370	0.0599	0.1458	0.3581	0.6765	0.8874	0.9853
i=4	0.1561	0.1359	0.0880	0.0492	0.0441	0.0889	0.2495	0.5508	0.8107	0.9561
i=5	0.2731	0.2516	0.1757	0.1051	0.0493	0.0455	0.1328	0.3602	0.6539	0.9021
i=6	0.4746	0.4515	0.3360	0.2109	0.1080	0.0485	0.0618	0.1860	0.4567	0.7760
i=7	0.6994	0.6775	0.5621	0.4216	0.2487	0.1138	0.0494	0.0763	0.2442	0.5689
i=8	0.8966	0.8746	0.8041	0.6816	0.4792	0.2769	0.1246	0.0517	0.0846	0.3114
i=9	0.9833	0.9769	0.9541	0.8971	0.7560	0.5457	0.2994	0.1240	0.0488	0.1108
i=10	0.9998	0.9990	0.9977	0.9876	0.9539	0.8183	0.5917	0.3164	0.1303	0.0549

APPENDIX C: Relative Frequencies of Letters in Latin Languages

Table C.1: Relative Letter Frequencies in 15 Latin Languages

Letter	English	French	German	Spanish	Portuguese	Esperanto	Italian	Turkish
a	8.17%	7.64%	6.52%	11.53%	14.63%	12.12%	11.75%	12.92%
b	1.49%	0.90%	1.89%	2.22%	1.04%	0.98%	0.93%	2.84%
c	2.78%	3.26%	2.73%	4.02%	3.88%	0.78%	4.50%	1.46%
d	4.25%	3.67%	5.08%	5.01%	4.99%	3.04%	3.74%	5.21%
e	12.70%	14.72%	16.40%	12.18%	12.57%	9.00%	11.79%	9.91%
f	2.23%	1.07%	1.66%	0.69%	1.02%	1.04%	1.15%	0.46%
g	2.02%	0.87%	3.01%	1.77%	1.30%	1.17%	1.64%	1.25%
h	6.09%	0.74%	4.58%	0.70%	0.78%	0.38%	0.64%	1.21%
i	6.97%	7.53%	6.55%	6.25%	6.19%	10.01%	10.14%	9.60%
j	0.15%	0.61%	0.27%	0.49%	0.40%	3.50%	0.01%	0.03%
k	0.77%	0.07%	1.42%	0.01%	0.02%	4.16%	0.01%	5.68%
l	4.03%	5.46%	3.44%	4.97%	2.78%	6.10%	6.51%	5.92%
m	2.41%	2.97%	2.53%	3.16%	4.74%	2.99%	2.51%	3.75%
n	6.75%	7.10%	9.78%	6.71%	4.45%	7.96%	6.88%	7.99%
o	7.51%	5.80%	2.59%	8.68%	9.74%	8.78%	9.83%	2.98%
p	1.93%	2.52%	0.67%	2.51%	2.52%	2.76%	3.06%	0.89%
q	0.10%	1.36%	0.02%	0.88%	1.20%	0	0.51%	0
r	5.99%	6.69%	7.00%	6.87%	6.53%	5.91%	6.37%	7.72%
s	6.33%	7.95%	7.27%	7.98%	6.81%	6.09%	4.98%	3.01%
t	9.06%	7.24%	6.15%	4.63%	4.34%	5.28%	5.62%	3.31%
u	2.76%	6.31%	4.17%	2.93%	3.64%	3.18%	3.01%	3.24%
v	0.98%	1.84%	0.85%	1.14%	1.58%	1.90%	2.10%	0.96%
w	2.36%	0.05%	1.92%	0.02%	0.04%	0	0.03%	0
x	0.15%	0.43%	0.03%	0.22%	0.25%	0	0.00%	0
y	1.97%	0.13%	0.04%	1.01%	0.01%	0	0.02%	3.34%
z	0.07%	0.33%	1.13%	0.47%	0.47%	0.49%	1.18%	1.50%
à	0	0.49%	0	0	0.07%	0	0.64%	0
â	0	0.05%	0	0	0.56%	0	0	0
á	0	0	0	0.50%	0.12%	0	0	0
å	0	0	0	0	0	0	0	0
ä	0	0	0.58%	0	0	0	0	0
ã	0	0	0	0	0.73%	0	0	0
a̧	0	0	0	0	0	0	0	0
æ	0	0	0	0	0	0	0	0
œ	0	0.02%	0	0	0	0	0	0
ç	0	0.09%	0	0	0.53%	0	0	1.16%
ĉ	0	0	0	0	0	0.66%	0	0
d́	0	0	0	0	0	0	0	0
č	0	0	0	0	0	0	0	0

Table C.2: Relative Letter Frequencies in 15 Latin Languages Continued

Letter	English	French	German	Spanish	Portuguese	Esperanto	Italian	Turkish
ď	0	0	0	0	0	0	0	0
ð	0	0	0	0	0	0	0	0
è	0	0.27%	0	0	0	0	0.26%	0
é	0	1.50%	0	0.43%	0.34%	0	0	0
ê	0	0.22%	0	0	0.45%	0	0	0
ë	0	0.01%	0	0	0	0	0	0
ę	0	0	0	0	0	0	0	0
ě	0	0	0	0	0	0	0	0
ĝ	0	0	0	0	0	0.69%	0	0
?	0	0	0	0	0	0	0	1.13%
ĥ	0	0	0	0	0	0.02%	0	0
î	0	0.05%	0	0	0	0	0	0
ì	0	0	0	0	0	0	-0.03%	0
í	0	0	0	0.73%	0.13%	0	0.03%	0
ï	0	0.01%	0	0	0	0	0	0
ı	0	0	0	0	0	0	0	5.11%
ĵ	0	0	0	0	0	0.06%	0	0
ł	0	0	0	0	0	0	0	0
ñ	0	0	0	0.31%	0	0	0	0
ń	0	0	0	0	0	0	0	0
ň	0	0	0	0	0	0	0	0
ò	0	0	0	0	0	0	0.00%	0
ö	0	0	0.44%	0	0	0	0	0.78%
ô	0	0.02%	0	0	0.64%	0	0	0
ó	0	0	0	0.83%	0.30%	0	0	0
õ	0	0	0	0	0.04%	0	0	0
ø	0	0	0	0	0	0	0	0
ř	0	0	0	0	0	0	0	0
ŝ	0	0	0	0	0	0.39%	0	0
ş	0	0	0	0	0	0	0	1.78%
ś	0	0	0	0	0	0	0	0
š	0	0	0	0	0	0	0	0
ß	0	0	0.31%	0	0	0	0	0
ť	0	0	0	0	0	0	0	0
þ	0	0	0	0	0	0	0	0
ù	0	0.06%	0	0	0	0	-0.17%	0
ú	0	0	0	0.17%	0.21%	0	0.17%	0
û	0	0.06%	0	0	0	0	0	0
ŭ	0	0	0	0	0	0.52%	0	0
ü	0	0	1.00%	0.01%	0.03%	0	0	1.85%
ů	0	0	0	0	0	0	0	0
ý	0	0	0	0	0	0	0	0
ź	0	0	0	0	0	0	0	0
ż	0	0	0	0	0	0	0	0
ž	0	0	0	0	0	0	0	0

Table C.3: Relative Letter Frequencies in 15 Latin Languages Continued

Letter	Swedish	Polish	Dutch	Danish	Icelandic	Finnish	Czech
a	9.38%	10.50%	7.49%	6.03%	10.11%	12.22%	8.42%
b	1.54%	1.74%	1.58%	2.00%	1.04%	0.28%	0.82%
c	1.49%	3.90%	1.24%	0.57%	0	0.28%	0.74%
d	4.70%	3.73%	5.93%	5.86%	1.58%	1.04%	3.48%
e	10.15%	7.35%	18.91%	15.45%	6.42%	7.97%	7.56%
f	2.03%	0.14%	0.81%	2.41%	3.01%	0.19%	0.08%
g	2.86%	1.73%	3.40%	4.08%	4.24%	0.39%	0.09%
h	2.09%	1.02%	2.38%	1.62%	1.87%	1.85%	1.36%
i	5.82%	8.33%	6.50%	6.00%	7.58%	10.82%	6.07%
j	0.61%	1.84%	1.46%	0.73%	1.14%	2.04%	1.43%
k	3.14%	2.75%	2.25%	3.40%	3.31%	4.97%	2.89%
l	5.28%	2.56%	3.57%	5.23%	4.53%	5.76%	3.80%
m	3.47%	2.52%	2.21%	3.24%	4.04%	3.20%	2.45%
n	8.54%	6.24%	10.03%	7.24%	7.71%	8.83%	6.47%
o	4.48%	6.67%	6.06%	4.64%	2.17%	5.61%	6.70%
p	1.84%	2.45%	1.57%	1.76%	0.79%	1.84%	1.91%
q	0.02%	0	0.01%	0.01%	0	0.01%	0.00%
r	8.43%	5.24%	6.41%	8.96%	8.58%	2.87%	4.80%
s	6.59%	5.22%	3.73%	5.81%	5.63%	7.86%	5.21%
t	7.69%	2.48%	6.79%	6.86%	4.95%	8.75%	5.73%
u	1.92%	2.06%	1.99%	1.98%	4.56%	5.01%	2.16%
v	2.42%	0.01%	2.85%	2.33%	2.44%	2.25%	5.34%
w	0.14%	5.81%	1.52%	0.07%	0	0.09%	0.02%
x	0.16%	0.00%	0.04%	0.03%	0.05%	0.03%	0.03%
y	0.71%	3.21%	0.04%	0.70%	0.90%	1.75%	1.04%
z	0.07%	4.85%	1.39%	0.03%	0	0.05%	1.50%
à	0	0	0	0	0	0	0
â	0	0	0	0	0	0	0
á	0	0	0	0	1.80%	0	0.87%
å	1.34%	0	0	1.19%	0	0.00%	0
ä	1.80%	0	0	0	0	3.58%	0
ã	0	0	0	0	0	0	0
ą	0	0.70%	0	0	0	0	0
æ	0	0	0	0.87%	0.87%	0	0
œ	0	0	0	0	0	0	0
ç	0	0	0	0	0	0	0
ĉ	0	0	0	0	0	0	0
ć	0	0.74%	0	0	0	0	0
č	0	0	0	0	0	0	0.46%
ď	0	0	0	0	0	0	0.02%
ð	0	0	0	0	4.39%	0	0
è	0	0	0	0	0	0	0

Table C.4: Relative Letter Frequencies in 15 Latin Languages Continued

Letter	Swedish	Polish	Dutch	Danish	Icelandic	Finnish	Czech
é	0	0	0	0	0.65%	0	0.63%
ê	0	0	0	0	0	0	0
ë	0	0	0	0	0	0	0
ȩ	0	1.04%	0	0	0	0	0
ě	0	0	0	0	0	0	1.22%
ĝ	0	0	0	0	0	0	0
ğ	0	0	0	0	0	0	0
ĥ	0	0	0	0	0	0	0
î	0	0	0	0	0	0	0
ì	0	0	0	0	0	0	0
í	0	0	0	0	1.57%	0	1.64%
ï	0	0	0	0	0	0	0
ı	0	0	0	0	0	0	0
ĵ	0	0	0	0	0	0	0
ł	0	2.11%	0	0	0	0	0
ñ	0	0	0	0	0	0	0
ń	0	0.36%	0	0	0	0	0
ň	0	0	0	0	0	0	0.01%
ò	0	0	0	0	0	0	0
ö	1.31%	0	0	0	0.78%	0.44%	0
ô	0	0	0	0	0	0	0
ó	0	1.14%	0	0	0.99%	0	0.02%
õ	0	0	0	0	0	0	0
ø	0	0	0	0.94%	0	0	0
ř	0	0	0	0	0	0	0.38%
ŝ	0	0	0	0	0	0	0
ş	0	0	0	0	0	0	0
ś	0	0.81%	0	0	0	0	0
š	0	0	0	0	0	0	0.69%
ß	0	0	0	0	0	0	0
ť	0	0	0	0	0	0	0.01%
þ	0	0	0	0	1.46%	0	0
ù	0	0	0	0	0	0	0
ú	0	0	0	0	0.61%	0	0.05%
û	0	0	0	0	0	0	0
ŭ	0	0	0	0	0	0	0
ü	0	0	0	0	0	0	0
ů	0	0	0	0	0	0	0.20%
ý	0	0	0	0	0.23%	0	1.00%
ź	0	0.08%	0	0	0	0	0
ż	0	0.71%	0	0	0	0	0
ž	0	0	0	0	0	0	0.72%