University of Khartoum
Faculty of Mathematical Sciences and Informatics
Department of Statistics

A graduation project for the B.Sc degree in Statistics
In:

Statistical distances using for constructing nonparametric tests

Prepared by:
1- Eisa Bush Ibrahim Manago.
2- Mariam Osman Hmed Idrees.
3- Mohamed Azeem Saeed Hajalamein.

Supervisor: Dr. Ibrahim Elabid

January 2025

Abstract

Statistical distances are regarded as key instruments in determining differences between points across multiple statistical fields, starting from measures in descriptive data represented by measures of dispersion. And they emerge in applications of simple linear regression This review highlights statistical distances and their role in constructing non-parametric tests such as the Chi-square test, which is distinguished by its usage of categorical data, as clear from the aforesaid study and shown in the COVID-19 results. Other than that, we find the K-nearest neighbors (K-NN) method and its vital role in machine learning, where it is employed in the classification and regression. The Kolmogorov-Smirnov test gives accuracy in comparing distributions and flexibility for multidimensional data, as in the ddks type. We find the Wasserstein distance and its relevance in inferential statistics, optimal transport, and the Cramér-von Mises distance.

Through this study, researchers can be prepared with the right criteria for selecting relevant tests, and this review affirms the essential role of statistical distances in non-parametric tests.

Introduction

Statistical distances are considered the cornerstone in the field of data analysis and statistical inference, as they assist in analyzing the similarities and differences between samples, aid in model estimation, and aid in the development of statistical tests. They are widely prevalent in statistical literature. However, the concepts associated with these distances still require a complete evaluation. This study seeks to present the notion of statistical distances and their importance in constructing non-parametric tests that do not require assumptions about the distribution of the data from which the sample is drawn. The value of this review is in creating a suitable infrastructure that lets scholars in the future readily access the idea of statistical distances and their role in non-parametric tests. This is achieved by delivering these tests and the corresponding tests for each type of data. The Chi-square test stands out in the analysis of categorical data, and the Kolmogorov-Smirnov test is chosen for comparing distributions in small samples. The Wasserstein distances provide a solution for comparing distributions in multi-dimensional spaces.

This paper is broken into three pieces. First, we will discuss statistical distances in measures of dispersion, starting with the simplest mathematical representation of statistical distances when determining the range. Then, we will move on to statistical distances and their role in simple linear regression, where they emerge in the residual analysis. The third and most essential component is statistical distances in non-parametric testing, which is our focus, as we will study numerous tests such as the Chi-square test for independence and the K-nearest neighbors (K-NN) method used in classification.And the

Kolmogorov-Smirnov test for comparing distributions and the Wasserstein and Cramér-von Mises distances.

Statistical Distances in Measures of Dispersion

It is also known as measures of variation, and it is used to describe data, such as the range, variance, and standard deviation.(Bluman, 2012) Statistical distances in these measures can be clearly found when we calculate each of these measures.

Range and Interquartile Range

The range is the difference between the maximum and minimum values.

$$\text{Range} = \text{maximum value} - \text{minimum value}$$

The Interquartile Range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1). The IQR represents the range of the middle half of the data

$$\text{IQR} = Q_3 - Q_1$$

.(Bluman, 2012)

Variance and Standard Deviation

The variance is the average of the squares of the distance each value is from the mean. The formula for the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X-\mu)^2$$

Where:

- X= individual value.
- μ = is the population mean.
- N= is the population size.

The standard deviation is the square root of the variance, and it indicates the average distance from the mean.

$$\sigma = \sqrt{\sigma^2}$$

.(Bluman, 2012)

Statistical Distances in Simple Regression

The simple linear regression model consists of the mean function and the variance function.

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

$$\text{Var}(Y|X = x) = \sigma^2$$

Statistical distances play a pivotal role in regression analysis; the method of least squares is used to estimate β_0 and β_1. We estimate β_0 and β_1 so that the sum of the squares of differences between the observations y_i and the straight line is minimized.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The difference between the observed value y_i and the fitted value $\hat{y}_i$ is a residual. Mathematically the i-th residual is

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \ldots, n$$

Residuals play an important role in investigating model adequacy and in detecting departures from underlying assumptions.(Montgomery et al., 2012)

Statistical distances in Nonparametric Tests

Nonparametric tests have several advantages over parametric tests: they are easier to use and understand; they can be applied to situations in which parametric tests cannot be used; and they do not require that the population being sampled is normally distributed. However, a major problem with nonparametric tests is that they are less efficient than parametric tests. The sample size must be larger for a nonparametric test to have the same probability of committing the two types of errors. (Weiss, 2012) Nonparametric tests are valid for data from any population, and may be used on nominal scale of measurement, or ordinal scale of data. They may also be used on data with an interval or ratio scale of measurement. (Conover,1999) Here we delve into statistical distances in nonparametric statistical tests like the Kolmogorov-Smirnov, Wasserstein, and Cramér-von Mises distances to measure dissimilarity between probability distributions.

Chi-square Test

The Chi-square is a significant statistic and should be followed with a strength statistic. Advantages of the chi-square include its robustness to the distribution of the data, its ease of computation, the detailed information that can be derived from the test, its flexibility

in handling data from both two-group and multiple-group studies. Limitations include its sample size requirements, the difficulty of interpretation when there are large numbers of categories (20 or more) in the independent or dependent variables. (McHugh,2013). The chi-square test for independence is also called Pearson's chi-square test or the chi-square test of association. It is used to discover if there is a relationship between two categorical variables (Boduszek,2016).
In the chi-square test for independence, the chi-square was evaluated in exploring the association between chronic disease, vaccination status, and recovery from COVID-19. The emphasis was on two cohorts: persons with chronic illnesses who remained unvaccinated and those without chronic conditions who were vaccinated. To comprehend the influence of factors on COVID-19 recovery outcomes. Upon gathering the requisite data from the patients, the results were presented in the table below.(Elsayed, 2022)

Table 1: Infection with Coronavirus (COVID-19)

Row Labels	**Death**	**Healing**	**Row marginals (Row sum)**
Free of chronic diseases and got the Vaccine	14	102	116
chronic disease and don't got the vaccine	176	24	200
Column marginals (Sum of the column)	190	126	$N = 316$

Statistical distances in computing chi-square

With the data in Table (1), the researcher can proceed with computing the χ^2 statistic to find out if the incidence of chronic diseases, not getting the vaccine, influences healing from non-healing (death). The formula for computing a Chi-Square is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i: Observed (the actual count of cases in each cell of the table).
- E_i: Expected value (estimated below).
- χ^2: The cell Chi-square value.

The formula evaluates how different the observed data is from the data we would expect on the assumption that there is no relationship. Here the statistical distance was shown in computation in $(O_i - E_i)$, and the interpretation is that if there is a small difference, then the two variables are independent, rather than a huge difference showing that the two variables are dependent. The primary result after computing the chi-square is shown in Table (2).(Elsayed, 2022)

Table 2: Cell expected values and (cell Chi-square values)

Expected (fe)	**Death**	**Healing**
Free of chronic diseases, got the vaccine	69.75 (44.6)	46.3 (67.2)
chronic disease, don't got the vaccine	120.3 (25.84)	79.75 (38.96)

It can be observed in Table (2) that the maximum cell χ^2 value of 67.2 occurs in Cell (2). This is a result of the measured value being 102, whereas only 46.3 were expected. Therefore, this cell has a substantially larger number of observed cases than would be expected by chance. Cell (2) reflects the second group: individuals with COVID-19 and free of chronic diseases. They have the same medical care and have been healed of COVID-19 (healing). This suggests that the quantity of the second group was much more than expected. The second greatest cell χ^2 value of 4.56 is located in Cell (1). However, in this cell, we see that the number of observed cases was substantially lower than expected (Observed = 44.6, Expected = 69.75). We note from the values of the table that there is a significant association between the healing of COVID-19 and chronic diseases.(Elsayed, 2022)

K-Nearest Neighbors (K-NN)

The k-nearest neighbor classifier assigns an instance to the class most heavily represented among its neighbors. It is based on the idea that the more similar the instances, the more likely it is that they belong to the same class. We can use the same approach for classification as long as we have a reasonable similarity or distance measure (Chen et al. 2009).

Given a training set Ξ of m labeled patterns, a nearest-neighbor procedure decides that some new pattern, $\mathbf{X}$, belongs to the same category as do its closest neighbors in Ξ. More precisely, a k-nearest-neighbor method assigns a new pattern, $\mathbf{X}$, to that category to which the plurality of its k closest neighbors belong. Using relatively large values of k decreases the chance that the decision will be unduly influenced by a noisy training pattern close to $\mathbf{X}$. But large values of k also reduce the acuity of the method. The k-nearest-neighbor method can be thought of as estimating the values of the probabilities of the classes given $\mathbf{X}$. Of course the denser are the points around $\mathbf{X}$, and the larger the value of k, the better the estimate.(Nilsson, 1998)

The distance metric used in nearest-neighbor methods (for numerical attributes) can be simple Euclidean distance. That is, the distance between two patterns $(x_{11}, x_{12}, \ldots, x_{1n})$ and $(x_{21}, x_{22}, \ldots, x_{2n})$ is $\sqrt{\sum_{j=1}^{n}(x_{1j} - x_{2j})^2}$. This distance measure is often modified by scaling the features so that the spread of attribute values along each dimension is approximately the same. In that case, the distance between the two vectors would be $\sqrt{\sum_{j=1}^{n} a_j^2(x_{1j} - x_{2j})^2}$, where a_j is the scale factor for dimension j. An example of a

nearest-neighbor decision problem is shown in Fig. 1 In the figure the class of a training pattern is indicated by the number next to it.(Nilsson, 1998)

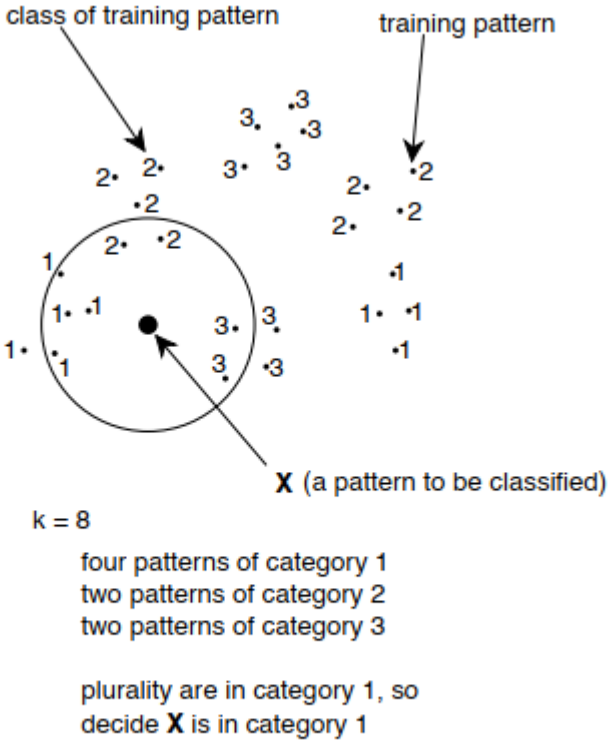

Figure 1: An 8-Nearest-Neighbor Decision

Nearest-neighbor methods are memory intensive because a large number of training patterns must be stored to achieve good generalization. Since memory cost is now reasonably low, the method and its derivatives have seen several practical applications. (Moore, 1992) and (Moore et al., 1994). Also, the distance calculations required to find nearest neighbors can often be efficiently computed by kd-tree methods (Friedman et al., 1977). A theorem by Cover and Hart (Cover & Hart, 1967) relates the performance of the 1-nearest-neighbor method to the performance of a minimum-probabilityof-error classifier.

Kolmogorov-Smirnov Test

Definition:
Let $X_1, X_2, \ldots, X_n$ be a random sample. The empirical distribution function, $S(x)$ is a function of x, which equals the fraction of X's that are less than or equal to x for each x, $-\infty < x < \infty$.(Conover,1999)

If we want to see if two or more samples are governed by the same unknown distribution, it seems natural to compare the empirical distribution functions of those samples to see if they look somewhat similar. To be precise, however, some measure of disparity between or among these functions is needed. Kolmogorov and Smirnov developed statistical procedures that use the maximum vertical distance between these functions as a measure of how well the functions resemble each other.

A test for goodness of fit usually involves examining a random sample from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact a known, specified function. That is, the null hypothesis specifies some distribution function $F^*(x)$, perhaps graphically as in Figure 2, or perhaps as a mathematical function that may be graphed. A random sample $X_1, X_2, \ldots, X_n$ is then drawn from some population and is compared with $F^*(x)$ in some way to see if it is reasonable to say that $F^*(x)$ is the true distribution function of the random sample.(Conover,1999) One logical way of comparing the random sample with $F^*(x)$ is by means of the empirical distribution function $S(x)$, which was defined as the fraction of X's that are less than or equal to x for each x, $-\infty < x < \infty$.The empirical distribution function $S(x)$ is useful as an estimator of $F(x)$, the unknown distribution function of the X's. So we can compare the empirical distribution function $S(x)$ with the hypothesized distribution function $F^*(x)$ to see if there is good agreement. If there is not good agreement, then we may reject the null hypothesis and conclude that the true but unknown distribution function, $F(x)$, is in fact not given by the function $F^*(x)$ in the null hypothesis..(Conover,1999) But what sort of test statistic can we use as a measure of the discrepancy between $S(x)$ and $F^*(x)$? One of the simplest measures imaginable is the largest distance between the two graphs $S(x)$ and $F^*(x)$, measured in a vertical direction. This is the statistic suggested by Kolmogorov (1933). That is, if $F^*(x)$ is given by Figure 2 and a random sample of size 5 is drawn from the population, the empirical distribution function $S(x)$ may be drawn on the same graph along with $F^*(x)$, as shown in Figure 3. If $F^*(x)$ and $S(x)$ are as given, the maximum vertical distance between the two graphs occurs just before the third step of $S(x)$. This distance is about 0.5 in Figure 2; therefore the Kolmogorov statistic T equals 0.5 in this case. Large values of T as determined by Table Quantiles of the Kolmogorov Test Statistic lead to rejection of $F^*(x)$ as a reasonable approximation to the unknown true distribution function $F(x)$.

The Kolmogorov test may be preferred over the chi-squared test for goodness of fit if the sample size is small; the Kolmogorov test is exact even for small samples..(Conover,1999)

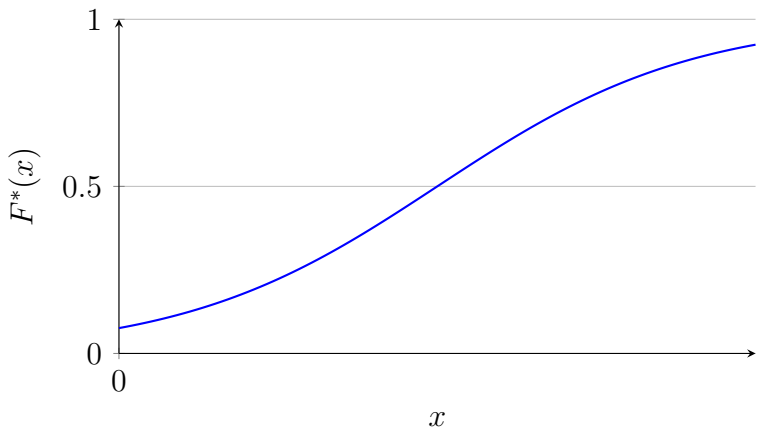

Figure 2: A hypothesized distribution function

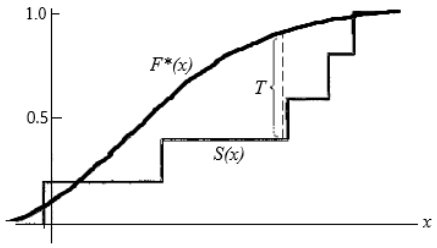

Figure 3: The hypothesized distribution function $F^*(x)$, the empirical distribution function $S(x)$, and Kolmogorov's statistic T.

Statistical distances in computing the Kolmogorov goodness-of-fit Test

Data: The data consist of a random sample $X_1, X_2, \ldots, X_n$ of size n associated with some unknown distribution function, denoted by $F(x)$.(Conover,1999)

Test Statistic: Let $S(x)$ be the empirical distribution function based on the random sample $X_1, X_2, \ldots, X_n$. The test statistic is defined differently for the three different sets of hypotheses A, B, and C. Let $F^*(x)$ be a completely specified hypothesized distribution function.(Conover,1999)
A. Two-Sided Test Let the test statistic T be the greatest (denoted by "sup" for supremum) vertical distance between $S(x)$ and $F^*(x)$. In symbols we say

$$T = \sup_x |F^*(x) - S(x)|$$

which is read "T equals the supremum, over all x, of the absolute value of the difference $F^*(x) - S(x)$".
B. One-Sided Test Denote this test statistic by T^+, and let it equal the greatest vertical distance attained by $F^*(x)$ above $S(x)$; that is,

$$T^+ = \sup_x [F^*(x) - S(x)]$$

which is similar to T except that we consider only the greatest difference where the function $F^*(x)$ is above the function $S(x)$
C. One-Sided Test For this test use the test statistic T^-, defined as the greatest vertical distance attained by $S(x)$ above $F^*(x)$. Formally this becomes

$$T^- = \sup_x [S(x) - F^*(x)]$$

DDKS Test Statistic

The d-dimensional extension of the Kolmogorov-Smirnov (ddKS) test statistic. (Fasano& Franceschini, 1987)The ddKS test statistic is the maximum of the differences between the cumulative distribution function (CDF) and survival function (SF) of two samples. We begin with two finite sets of samples in $\mathbb{R}^d$ which we denote as P and T. We first assume that the Cartesian coordinate system is an appropriate basis for the samples provided. Then, we can construct $2d$ cumulative density estimates, where the cumulative density is defined as the number of points existing in an orthant relative to a given point. This is equivalent in two dimensions to counting the number of points in the top-right, bottom-right, bottom-left, and top-left quadrants relative to a chosen location. Given these two cumulative density estimates $C_P(\vec{x})$ and $C_T(\vec{x})$, the ddKS test statistic is simply the maximum absolute difference between the two.(Hagen et al., 2021)

$$D = \max |C_P(\vec{x}) - C_T(\vec{x})|$$

Hagen et al.'s result is that ddKS works well on all datasets, whereas each other technique has poor power on at least one of the studied datasets, and that ddKS is very capable

of detecting distribution differences in distributions contaminated by noise. They also explored the behavior of ddKS's accelerated approximations, proving that these are good approximations to ddKS itself and that they are, in fact, quicker. While ddKS has a time complexity of $O(2^d N^2)$, vdKS has a time complexity of $O(2^d Nk)$ and rdKS has a time complexity of $O((d+1)N \log N)$. These time complexities are small enough for use in most physical science applications, and, while still restrictive for the very high-dimensional applications present in machine learning fields, represent a significant step towards computationally tractable high-dimensional non-parametric test statistic calculation.(Hagen et al., 2021)

The Wasserstein distances

Wasserstein distances are metrics between probability distributions that are inspired by the problem of efficient transportation. These distances (and the optimal transport problem) are ubiquitous in mathematics, particularly in fluid mechanics, partial differential equations, optimization, and, of course, probability. In addition to their theoretical importance, they have provided a successful framework for the comparison of (at times complex) objects in fields of application such as image retrieval (Rubner et al., 2000), computer vision (Ni et al., 2009), pharmaceutical statistics (Munk and Czado, 1998), genomics (Bolstad et al., 2003; Evans and Matsen, 2012), economics (Gini, 1914), and finance (Rachev et al., 2011), to name but a few. Indeed, while their origins lie with Monge's (primarily mathematical) enquiry into how to optimally transport a pile of earth of a given volume into a pit of equal volume but potentially different shape, assume that we are in charge of the transport of goods between producers and consumers, whose respective spatial distributions are modeled by probability measures. The more manufacturers and consumers are far away from one another, the more difficult our job will be, and we would like to summarize the degree of difficulty with only one quantity. The best transit cost between the two measures:

$$C(\mu;\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\, d\pi(x,y), \tag{1}$$

where $c(x,y)$ is the cost for transporting one unit of mass from x to y. Here we do not care about the shape of the optimizer but only in the value of this optimal cost. One can think of (1) as a kind of distance between μ and ν, but in general it does not, strictly speaking, satisfy the axioms of a distance function. However, when the cost is defined in terms of a distance, it is easy to cook up a distance from (1).(Villani, 2008)

Definition: Wasserstein distances

Let $(\mathcal{X}, d)$ be a Polish metric space, and let $p \in [1, \infty)$. For any two probability measures

μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x,y)^p \, d\pi(x,y) \right)^{1/p},$$

or

$$= \inf \left\{ (E[d(X,Y)^p])^{1/p} : \text{law}(X) = \mu, \text{law}(Y) = \nu \right\}. \tag{2}$$

The distance W_1 is also called the Kantorovich-Rubinstein distance.(Villani, 2008)
Definition: Wasserstein space
With the same conventions as in definition of Wasserstein distances, the Wasserstein space of order p is defined as

$$P_p(\mathcal{X}) := \{\mu \in P(\mathcal{X}) : \int_{\mathcal{X}} d(x_0,x)^p \, \mu(dx) < +\infty\},$$

where $x_0 \in \mathcal{X}$ is arbitrary. This space does not depend on the choice of the point x_0. Then W_p defines a (finite) distance on $P_p(\mathcal{X})$. In words, the Wasserstein space is the space of probability measures which have a finite moment of order p. In these notes, it will always be equipped with the distance W_p.

Proof. Proof that W_p is finite on P_p. Let π be a transference plan between two elements μ and ν in $P_p(\mathcal{X})$. Then the inequality

$$d(x,y)^p \leq 2^{p-1} \left[d(x,x_0)^p + d(x_0,y)^p \right]$$

shows that $d(x,y)^p$ is $\pi(dx\, dy)$-integrable as soon as $d(\cdot,x_0)^p$ is μ-integrable and $d(x_0,\cdot)^p$ is ν-integrable. □ □

Wasserstein distances appear in statistics in several ways. We delineate three broad categories of statistical use of these distances, according to which we will structure our review:

Optimal Transport as a Technical Tool

Wasserstein distances and the associated notion of an optimal coupling are often exploited as a versatile tool in asymptotic theory, due to the topological structure they induce and their relatively easy majorisation, and this section reviews some of the features of Wasserstein metrics that make them useful as a technical tool for deriving large sample theory results in statistics. To facilitate the presentation, we first state some simple facts that play a role in the development. Let X and Y be random vectors taking values in $\mathcal{X} = \mathbb{R}^d$; we maintain the notation $(\mathcal{X}, \|\cdot\|)$ to stress that the properties are valid in infinite dimensions as well.

- For any real number a, $W_p(aX, aY) = |a|W_p(X, Y)$.
- For any fixed vector $x \in \mathcal{X}$, $W_p(X + x, Y + x) = W_p(X, Y)$.
- For any fixed $x \in \mathcal{X}$, we have $W_2^2(X+x, Y) = \|x+\mathbb{E}(X)-\mathbb{E}(Y)\|^2+W_2^2(X, Y)$.
- For product measures and when $p = 2$, we have $W_2^2(\otimes_{i=1}^n \mu_i, \otimes_{i=1}^n \nu_i) = \sum_{i=1}^n W_2^2(\mu_i, \nu_i)$ in the analytic notation.

The proofs of the first three statements rely on the equivalence between the classes of the corresponding couplings. For example, $U = (X + x, Y + x)$ is a coupling of $X + x$ and $Y + y$ if and only if $U - (x, x)$ is a coupling of (X, Y). For the last property, observe that the map:

$$x \mapsto \left[t_{\nu_1}^{\mu_1}(x), \ldots, t_{\nu_1}^{\mu_1}(x)\right]$$

is a gradient of a convex function and pushes forward $\otimes\mu_i$ to $\otimes\nu_i$.(Villani, 2008)

Equilibrium, Concentration, and Poisson Approximations

A different class of settings where Wasserstein distances are used is in the study of convergence of Markov chains to their equilibrium distribution and dates back to Dobrushin (1970). The idea is to show a sort of contraction property of the transition kernel with respect to the Wasserstein distance. Let P be the transition matrix. In studying convergence of the Kac random walk on the orthogonal group $SO(n)$, Oliveira (2009) showed that

$$W_{D,2}(\mu P, \nu P) \leq \xi W_{D,2}(\mu, \nu)$$

for some $\xi < 1$, where D is a distance between matrices, leading to exponential convergence to equilibrium. A result of similar spirit is derived by Eberle (2014) for the transition kernel of the Metropolis adjusted Langevin algorithm, a Markov chain Monte Carlo method. The constant ξ above is related to the Wasserstein spectral gap of the transition kernel. Hairer et al. (2014) explore its behavior in infinite-dimensional state spaces, when taking finite-dimensional projections of P. They show that for the preconditioned Crank-Nicolson algorithm, ξ remains stable, whereas for the random walk Metropolis algorithm, ξ may converge to one. Rudolf & Schweizer (2018) employ Wasserstein distances to bound the difference between the behavior of some "nicely behaved" Markov chain and a perturbed version thereof, obtained from a modification in the transition kernel.

Optimal Transport as a Tool for Inference

As a measure of distance between probability laws, the Wasserstein distance can be used for carrying out goodness-of-fit tests, and indeed this has been its main use as a

tool for statistical inference. In the simplest one-sample setup, we are given a sample $X_1, \ldots, X_n$ with unknown law μ and wish to test whether μ equals some known fixed law μ_0 (e.g., standard normal or uniform). The empirical measure μ_n associated with the sample $(X_1, \ldots, X_n)$ is the (random) discrete measure that assigns mass $1/n$ to each observation X_i. In this sense, the strong law of large numbers holds in Wasserstein space: with probability one, $W_p(\mu_n, \mu) \to 0$ as $n \to \infty$ if and only if $\mathbb{E}\|X\|^p < \infty$. It is consequently appealing to use $W_p(\mu_n, \mu_0)$ as a test statistic. In the two-sample setup, one independently observes a sample $Y_1, \ldots, Y_m \sim \nu$ with corresponding empirical measure ν_m, and $W_p(\mu_n, \nu_m)$ is a sensible test statistic for the null hypothesis $\mu = \nu$.(Villani, 2008)

Cramér-von Mises distances

Baringhaus and Henze (2017) give a probabilistic interpretation of the Cramér–von Mises distance

$$\Delta(F, F_0) = \int (F - F_0)^2 \, dF_0$$

between continuous distribution functions F and F_0. If F is unknown, they construct an asymptotic confidence interval for $\Delta(F, F_0)$ based on a random sample from F. Moreover, for a given F_0 and some value $\Delta_0 > 0$, they propose an asymptotic equivalence test of the hypothesis that $\Delta(F, F_0) \geq \Delta_0$ against the alternative $\Delta(F, F_0) < \Delta_0$. If such a 'neighborhood-of-F_0 validation test,' carried out at a small asymptotic level, rejects the hypothesis, there is evidence that F is within a distance Δ_0 of F_0. As a neighborhood-of-exponentiality test shows, the method may be extended to the case that H_0 is composite. The Cramér–von Mises distance

$$\Delta(F, F_0) = \int_{-\infty}^{\infty} (F(x) - F_0(x))^2 \, dF_0(x)$$

between continuous distribution functions is one of the distinguished measures of deviation between distributions.(Baringhaus & Henze, 2017)
$\Delta(F, F_0)$ seems to be difficult to interpret, but they show that $\frac{2}{3} + \Delta(F, F_0)$ can be regarded as a probability.
The cramer-von Mises distance is a nonparametric measure used in hypothesis testing to assess the difference between empirical distribution function. It quantifies the discrepancy between distributions, facilitating the detection of change points in sequences of independent obseravtions. (Erlemann, 2021)
The cramer-von Mises distance is applied to the distribution of the excess over a confidence level and a new confidence interval for the related fitting error. (Gaigall & Gerstenberg, 2023)
The cramer-von Mises distances are used for comparing distributions, assessing whether

a sample stems from a predefined continuous distribution , and determining if two discrete random vector samples originate from the same underlying distribution. (Hanebeck, 2008)

Discussion

This review has shown that statistical distances are excellent tools to construct nonparametric tests, and the important conclusions are as follows:
According to Elsayed (2022), the Chi-square test is adequate for examining frequency tables, but its limitations include its sample size requirements and the difficulty of interpretation when there are large numbers of categories (20 or more) in the independent or dependent variables. This conclusion is based on a comparison of statistical distances in nonparametric tests. The K-Nearest Neighbors method, which uses Euclidean distance, has some limitations.The disadvantage is that it requires the entire training set to be saved, which increases the storage needs.
The Kolmogorov-Smirnov test is known for its accuracy, particularly in small sample sizes. However, its calculation method is based on the maximum distance between distributions, making it sensitive to extreme values. The Wasserstein distance is a strong tool for comparing distributions in multidimensional spaces; nevertheless, its computational complexity requires more time, limiting its utility in multidimensional spaces.

Conclusion

Statistical distances are important for developing and improving nonparametric tests. This form of test does not require any prior assumptions about data distribution, making it an extremely useful analytical tool. This review indicated that the relevant statistical distances depend on the type of data under study.
Furthermore, this review suggests that future research should focus on building algorithms to aid in the conduct and power of non-parametric tests. It also proposes performing more reviews like this one to investigate the benefits and create a combination between statistical distances, thus enhancing the power of non-parametric tests.
Based on what has already been discussed, this review represents a useful framework for selecting and using statistical distances.

References

1. Baringhaus, L., & Henze, N. (2017). Cramér–von Mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *Journal of Nonparametric Statistics*, 29(2), 167–188.

2. Bluman, A. G. (2012). Bluman, Elementary Statistics: A Step by Step Approach, © 2009, 8e, Student Edition (Reinforced Binding) with Formula Card. McGraw-Hill Education.

3. Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), 185-193.

4. Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 11, 747-776.

5. Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Wiley.

6. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21-27.

7. Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability and Its Applications*, 15(3), 458-486.

8. Eberle, A. (2014). Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Annals of Probability*, 24(2), 337-377.

9. Elsayed, K. A. (2022). Chi-square independence test: To study the effect of chronic diseases and recovery from COVID-19. Case study King Abdullah Hospital during the period December 2020-June 2021. *Al-Magalla Al-Miṣriyya Lil Dirasat Al-Tigariyya (Online)*, 46(4), 306-335.

10. Erlemann, R. (2021). Cramér-von Mises tests for change points. *Scandinavian Journal of Statistics*, 49(2), 802-830.

11. Evans, S. N., & Matsen, F. A. (2012). The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4), 569-592.

12. Fasano, G., & Franceschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225, 155-170.

13. Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3), 209-226.

14. Gaigall, D., & Gerstenberg, J. (2023). Cramér-von-Mises tests for the distribution of the excess over a confidence level. *Journal of Nonparametric Statistics*, 1–33.

15. Gini, C. (1914). Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 74, 185-213.

16. Hagen, A., Jackson, S., Kahn, J., Strube, J. F., Haide, I., Pazdernik, K., & Hainje, C. (2021). Accelerated computation of a high-dimensional Kolmogorov-Smirnov distance. arXiv: Computation.

17. Hanebeck, U. D., & Kliesch, V. (2008). Localized Cumulative Distributions and a Multivariate Generalization of the Cramer-von Mises

18. Hairer, M., Stuart, A. M., & Vollmer, S. J. (2014). Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Annals of Applied Probability*, 24(6), 2455-2490.

19. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 30-37.

20. McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 23(2), 143-149.

21. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

22. Moore, A. (1992). Fast, robust adaptive control by learning only forward models. In J. Moody, S. Hanson, & R. Lippman (Eds.), *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann.

23. Moore, A. W., Hill, D. J., & Johnson, M. P. (1994). An empirical investigation of brute force to choose features, smoothers, and function approximators. In S. Hanson, S. Judd, & T. Petsche (Eds.), *Computational Learning Theory and Natural Learning Systems* (Vol. 3). MIT Press.

24. Munk, A., & Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2), 223-241.

25. Ni, K., Bresson, X., Chan, T., & Esedoglu, S. (2009). Local histogram-based segmentation using the Wasserstein distance. *International Journal of Computer Vision*, 84(1), 97-111.

26. Nilsson, N. J. (1998). *Introduction to Machine Learning: An Early Draft of a Proposed Textbook*.

27. Oliveira, R. I. (2009). On the convergence to equilibrium of Kac's random walk on matrices. *Annals of Applied Probability*, 19(3), 1200-1231.

28. Rachev, S. T., Stoyanov, S. V., & Fabozzi, F. J. (2011). *A Probability Metrics Approach to Financial Risk Measures*. Wiley.

29. Rippl, T., Munk, A., & Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151, 90-109.

30. Rudolf, D., & Schweizer, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4), 2610-2639.

31. Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99-121.

32. Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

33. Weiss, N. A. (2012). *Introductory Statistics* (9th ed.). Pearson.