Representativeness Testing for Samples on Graphs

Freek

February 6, 2025

Click here for the latest version

Abstract

I develop a randomization test of a sample's population representativeness on induced graph test statistics for situations of absent unit-level characteristic data. I prove it is an exact level-α test. I show how demographic selection into the sample can lead to rejection of the test. I also show how a technique called convex hull peeling can be used to extend to a joint version of this test. I then apply the technique to social network and ride-hailing survey case studies.

1. Introduction

Economists often[1] conduct surveys on a sample of n units from a population of N units, with $n < N$. You can collect data on the characteristics of the n units you select for your survey. However, there are situations in which you are not able to observe characteristics of all remaining $N - n$ units in the population. In that case, you can not conduct standard t-tests on characteristic variables to test whether the selected sample is representative of the population, nor weight units in your sample to improve the representativeness of results dependent on your sample. Nevertheless, you then sometimes do observe a graph on the population of all N units.

I propose a hypothesis test for the representativeness of the sample that compares a test statistic computed from the induced subgraph for the selected n units to the same test statistic computed from a collection of induced subgraphs from random samples of n units. Examples of types of graphs this test can be applied to include the follower and friend network when studying a social medium. The test I propose to evaluate the population representativeness of the selected sample will reject the hypothesis when observing extreme graph statistics of the sample that selects into the survey relative to randomly selected samples that could select into the survey. A sample featuring extreme graph statistics can indicate the people in your sample have similar non-representative interests, or have similar non-representative demographics, as I show in the application section of this paper (section 3).

I apply the test to a randomized anti-conflict workshop allocation program among student networks in New Jersey schools. I reject the hypothesis that there is random assignment of the treatment. I then turn to a ride-hailing rider survey with simulated riding histories and confirm the test rejects the hypothesis that selection into the survey was random when riders from lower-income Census block groups are more likely to complete the survey.

The main limitations of the test statistic I propose are that it is not necessarily connected to demographic representativeness, and may be time-consuming to implement for researchers.

Section 2 presents the test and a derivation of its properties. Section 3 applies the test to a social network and road network scenario.

[1] See the multitude of results after searching for research based on the keyword "survey" on the AER website.

2. Graph Representativeness Test

2.1. Setup

Suppose the test statistic is permutation invariant. Define the test statistic as the surjective mapping T with $T : \mathcal{S} \mapsto \mathcal{T}$, where $\mathcal{S}$ is the domain of induced subgraphs of order n on the population graph G. The graph s having "order" n means it has n vertices, or $|V(s)| = n$, with $V(s)$ denoting the vertex set of s. Also, let the test statistic take real values, such that $\mathcal{T} \subset \mathbb{R}$.

Given this setup, there are at most $\binom{N}{n}$ unique values that $T(s)$ can take[2]. Denote this upper bound by u. The number of unique values that $T(s)$ takes is $|\mathcal{T}|$, where $|\mathcal{T}| \leq u < \infty$. With that, we can write $\mathcal{T}$ (given surjectivity of the map T) as $\mathcal{T} = \left\{t_1, t_2, \ldots t_{|\mathcal{T}|}\right\}$, with $t_k < t_l$ for all $k, l \in [|\mathcal{T}|]$, $k < l$.

Let X be the random variable taking the value of the test statistic T on the induced subgraph s, with s being constructed from a random sample of n units (sampled without replacement). Then we can write the probability that X is equal to any possible value of the test statistic t_k as

$$\mathbb{P}(X = t_k) = \frac{|\{s \in \mathcal{S} \ : \ T(s) = t_k\}|}{|\mathcal{T}|}. \tag{1}$$

Furthermore, let s_0 denote the induced subgraph from the sample of n units that selected into our study. We can then define the hypothesis that our sample is randomly drawn from the population as

$$H_0 : s_0 \overset{d}{=} X. \tag{2}$$

The test is structured as described in Algorithm 1.

Algorithm 1 Conducting the graph representativeness test

▷ *Set up the necessary values* ◁
Compute the test statistic $T(s_0)$ of your sample s_0
Define the sample size $n \leftarrow |s_0|$
Define the number of tests ν
for all $i \in [\nu]$ **do**
    Randomly draw without replacement the sample s_i of all N units, $|s_i| = n$
    Compute the test statistic $T(s_i)$

▷ *Determine the rank of the test statistic for the selected sample* ◁
Let $l \leftarrow \{i \in [\nu] : T(s_i) < T(s_0)\}$
Let $g \leftarrow \{i \in [\nu] : T(s_i) > T(s_0)\}$
Let $e \leftarrow \{i \in [\nu] : T(s_i) = T(s_0)\}$

▷ *Break ties* ◁
Define the ranks $\mathbf{r} = \{1, 2, \ldots, |e|+1\}$
for all $i \in [\nu]$ **do**
    Draw a "rank" r_i from Unif$[\mathbf{r}]$
    Update $\mathbf{r} \leftarrow \mathbf{r} \setminus \{r_i\}$

[2] If the order of vertices in s *does* matter for the value of $T(s)$, then there would be at most ${}_N P_n$ values that $T(s)$ can take, which is also a finite number.

Let $e_l \leftarrow \{i \in e : r_i < r_0\}$
Let $e_g \leftarrow \{i \in e : r_i > r_0\}$

▷ *Compute the p-value* ◁

Reject H_0 if $2\min\left\{\frac{|l|+|e_l|}{\nu}, \frac{|g|+|e_g|}{\nu}\right\} \leq \alpha$

2.2. Exactness Level of Test

I will now prove that the algorithm conducts an asymptotically exact level-α test.

First define the proportion of sample indices in $[\nu]$ for which the corresponding test statistic is equal to t_k out of all indices as $\hat{\mathbb{P}}(t_k) = \frac{|\{i \in [\nu] \,:\, T(s_i) = t_k\}|}{\nu}$.

Lemma 1. $\hat{\mathbb{P}}(t_k) \xrightarrow{a.s.} \mathbb{P}(X = t_k)$ *as* $\nu \to \infty$.

Proof. We can simplify the empirical probability as follows.

$$\begin{aligned}\hat{\mathbb{P}}(t_k) &= \frac{|\{i \in [\nu] \;:\; T(s_i) = t_k\}|}{\nu} \\ &= \frac{\sum_{i=1}^{\nu} \mathbb{1}(T(s_i) = t_k)}{\nu} \\ &= \frac{1}{\nu}\sum_{i=1}^{\nu} \mathbb{1}(T(s_i) = t_k)\end{aligned} \tag{3}$$

Because of the finite codomain of T and i.i.d. relationship between the randomly selected samples, by the strong law of large numbers $\hat{\mathbb{P}}(t_k) \xrightarrow{a.s.} \mathbb{E}[\mathbb{1}(T(s_i) = t_k)] = \mathbb{P}(X = t_k)$ as $\nu \to \infty$. ∎

Now, define the CDF of the random variable X by

$$\begin{aligned}F(\tau) &= \mathbb{P}(X \leq \tau) \\ &= \sum_{t \in \mathcal{T}} \mathbb{1}(t < \tau)\,\mathbb{P}(X = t)\,.\end{aligned} \tag{4}$$

Empirically, we may not have discovered all values the test statistic takes in the population. Define the set of values of the test statistic encountered in the ν samples as $\tilde{\mathcal{T}} \subseteq \mathcal{T}$. With this, I define the empirical CDF as $\hat{F}(\tau) = \sum_{t \in \tilde{\mathcal{T}}} \mathbb{1}(t < \tau)\,\hat{\mathbb{P}}(X = t)$.

Lemma 2. $\hat{F}(\tau) \xrightarrow{a.s.} F(\tau)$ *as* $\nu \to \infty$.

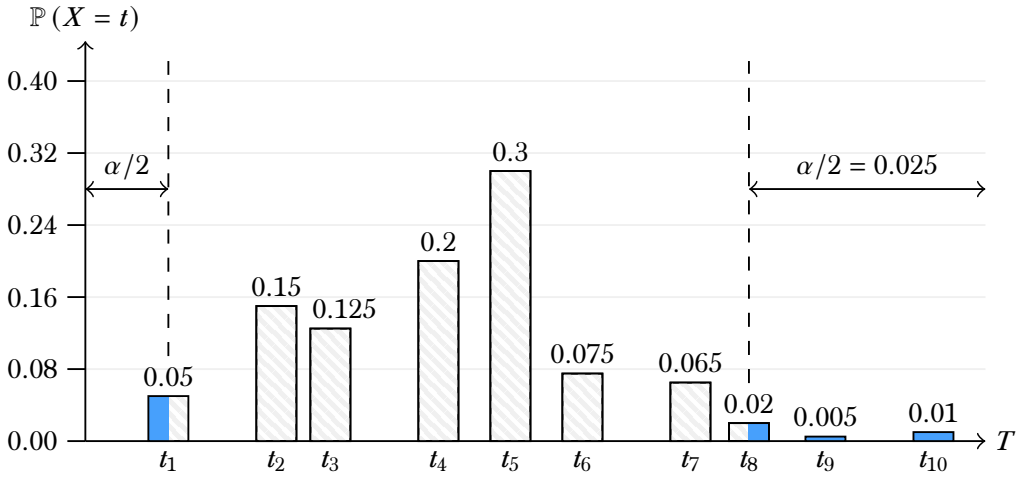
Proof. First, note that

$$\begin{aligned}\hat{F}(\tau) &= \sum_{t \in \tilde{\mathcal{T}}} \mathbb{1}(t < \tau)\,\hat{\mathbb{P}}(X = t) \\ &= \sum_{t \in \tilde{\mathcal{T}}} \mathbb{1}(t < \tau)\,\hat{\mathbb{P}}(X = t) + \sum_{t \in \mathcal{T} \setminus \tilde{\mathcal{T}}} \mathbb{1}(t < \tau)\underbrace{\hat{\mathbb{P}}(X = t)}_{=\,0} \\ &= \sum_{t \in \mathcal{T}} \mathbb{1}(t < \tau)\,\hat{\mathbb{P}}(X = t)\,.\end{aligned} \tag{5}$$

Applying the result from Lemma 1, the property of almost sure convergence that $Y_n \xrightarrow{a.s.} y$ and $Z_n \xrightarrow{a.s.} z$ implies $aY_n + bY_n \xrightarrow{a.s.} ay + bz$ to equation (5), and the fact that $|\mathcal{T}|$ is finite and fixed, we obtain that $\hat{F}(\tau) = \sum_{t\in\mathcal{T}} \mathbb{1}(t < \tau)\hat{\mathbb{P}}(X = t) \xrightarrow{a.s.} \sum_{t\in\mathcal{T}} \mathbb{1}(t < \tau)\mathbb{P}(X = t) = F(\tau)$. ■

With these definitions and results in mind, the proof the test is an asymptotically exact level-α test will proceed as illustrated in Figure 1 below.

Figure 1: Graphical Illustration of Proof



Notes: This figure shows the asymptotic probabilities $\mathbb{P}(X = t_k)$ for all t_k in the population, $\mathcal{T} = \{t_1, t_2, \ldots, t_{10}\}$. The asymptotic probabilities add up to one. If the test statistic of the selected sample t_0 takes a value associated with a fully non-blue bar, the test will asymptotically not be rejected that the sample is randomly selected from the population. If t_0 takes a value associated with a bar with some blue in it, the test will asymptotically be rejected with the probability equal to the proportional coverage of blue strictly within the corresponding bar. If the null hypothesis holds, we will asymptotically reject it with a probability equal to the proportion of space within the bars that is colored blue. In the picture, that proportion is 0.05.

Theorem 1. *Algorithm 1 rejects H_0 with probability α under the null as $\nu \to \infty$ with fixed $n, N \in \mathbb{N}$, $n \leq N$.*

Proof. Take α to be the level of the two-sided test, with $\alpha \in (0,1)$. Then there are two cases to consider in the proof that the test is at the level of α as $\nu \to \infty$: (i) $\exists t^* \in \mathcal{T} \ : \ F(t^*) = \alpha/2$, and (ii) $\nexists t^* \in \mathcal{T} \ : \ F(t^*) = \alpha/2$.

I will deal with case (i) first.

(a) If $t_0 < t^*$, then $\frac{|l|+|e_l|}{\nu} \leq \frac{\sum_{i=1}^{\nu} \mathbb{1}(T(s_i)<t^*)}{\nu} = \hat{F}(t^*) \xrightarrow{a.s.} F(t^*) = \alpha/2$ as $\nu \to \infty$.
So $\mathbb{P}\big(\text{reject } H_0 \text{ on left tail} \mid t_0 < t^*, \text{ (i)}\big) \xrightarrow{a.s.} 1$ as $\nu \to \infty$.

(b) If $t_0 > t^*$, then $\frac{|l|+|e_l|}{\nu} \geq \frac{|l|}{\nu} = \hat{F}(t_0) \xrightarrow{a.s.} F(t_0) > F(t^*) = \alpha/2$ as $\nu \to \infty$.
So $\mathbb{P}\big(\text{reject } H_0 \text{ on left tail} \mid t_0 > t^*, \text{ (i)}\big) \xrightarrow{a.s.} 0$ as $\nu \to \infty$.

(c) Suppose $t_0 = t^*$. $\mathbb{P}\left(\frac{|l|+|e_l|}{\nu} > \frac{|l|}{\nu}\right) = \mathbb{P}(|e_l| > 0) \xrightarrow{a.s.} 1$ because $\frac{|e|}{\nu} \xrightarrow{a.s.} \mathbb{P}(X = t_0) > 0$[3] implies

[3]Because T is a surjective map, there must be at least one induced subgraph s from n units in $\mathcal{S}$ such that $T(S) = t^*$. As $u = |\mathcal{S}| < \infty$, $\mathbb{P}(X = T(s)) \geq \frac{1}{|\mathcal{S}|} > 0$, with $|\mathcal{S}|$ being fixed, independent of ν.

that $|e|$ is $O_p(\nu)$ through satisfying[4] the almost sure convergence of $\frac{1}{\nu}\sum_{i=1}^{\nu}\mathbb{1}(T(s_i)=t_0)$ to $\mathbb{P}(X=t_0)$. We can thus derive the probability corresponding to the case at hand as shown in the below.

$$\begin{aligned}\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t^*,\ \text{(i)}\right) &= 1\cdot\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t^*,\ |e_l|=0,\ \text{(i)}\right)+\\ &\quad 0\cdot\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t^*,\ |e_l|>0,\ \text{(i)}\right)\\ &\xrightarrow{a.s.} 1\cdot 0+0\cdot 0\\ &= 0\end{aligned} \tag{6}$$

So $\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t^*,\ \text{(i)}\right)\xrightarrow{a.s.}0$ as $\nu\to\infty$.

Therefore, under H_0, by a property of almost sure convergence we have that

$$\begin{aligned}\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid \text{(i)}\right) &= \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0<t^*,\ \text{(i)}\right)\mathbb{P}(t_0<t^*)+\\ &\quad \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0>t^*,\ \text{(i)}\right)\mathbb{P}(t_0>t^*)+\\ &\quad \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t^*,\ \text{(i)}\right)\mathbb{P}(t_0=t^*)\\ &\xrightarrow{a.s.} 1\cdot\alpha/2+0\cdot(1-\alpha/2-\mathbb{P}(X=t^*))+0\cdot\mathbb{P}(X=t^*)\\ &= \alpha/2.\end{aligned} \tag{7}$$

In case (ii), consider the index $b\in[\mathcal{T}]$ such that $t_b=\min\{t\in\mathcal{T}\ :\ F(t)>\alpha/2\}$. Remember that without loss of generality, $t_k<t_l$ for all $k,l\in[|\mathcal{T}|]$, $k<l$. Consider $\rho:=\alpha/2-F(t_{b-1})$. t_{b-1} must exist because $F(t_b)=\mathbb{P}(X<t_b)>\alpha/2>0$. Note that $0<\rho<\mathbb{P}(X=t_{b-1})$. As defined earlier, e_l is the set of indices for which the test statistic value is equal to $T(s_0)$.

$$\begin{aligned}\mathbb{P}\left(\frac{|e_l|}{|e|}<\rho\right) &= \mathbb{P}(|e_l|<\rho|e|)\\ &= \mathbb{P}(|\{i\in e\ :\ r_i<r_0\}|<\rho|e|)\\ &= \frac{\lceil\rho|e|\rceil}{|e|+1}\end{aligned} \tag{8}$$

With $\nu\to\infty$, $\frac{\lceil\rho|e|\rceil}{|e|+1}=\frac{\frac{1}{\nu}\lceil\rho|e|\rceil}{\frac{1}{\nu}(|e|+1)}\xrightarrow{a.s.}\frac{\rho\mathbb{P}(X=t_0)}{\mathbb{P}(X=t_0)}=\rho$ by the algebra of limits and the sandwich theorem.

Therefore, consider the following cases:

(a) If $t_0<t_{b-1}$, then $\frac{|l|+|e_l|}{\nu}\le\frac{\sum_{i=1}^{\nu}\mathbb{1}(T(s_i)<t_{b-1})}{\nu}=\hat{F}(t_{b-1})\xrightarrow{a.s.}F(t_{b-1})<\alpha/2$ as $\nu\to\infty$.
So $\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0<t_{b-1},\ \text{(ii)}\right)\xrightarrow{a.s.}1$ as $\nu\to\infty$.

(b) If $t_0>t_{b-1}$, then $\frac{|l|+|e_l|}{\nu}\ge\frac{|l|}{\nu}=\hat{F}(t_0)\xrightarrow{a.s.}F(t_0)\ge F(t_b)>\alpha/2$ as $\nu\to\infty$.
So $\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0>t_{b-1},\ \text{(ii)}\right)\xrightarrow{a.s.}0$ as $\nu\to\infty$.

(c) $\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0=t_{b-1},\ \text{(ii)}\right)=\mathbb{P}\left(\frac{|e_l|}{|e|}<\rho\right)\xrightarrow{a.s.}\rho$ as shown in equation (8).

[4] This is, in turn, implied by satisfying the almost sure convergence of $\frac{1}{\nu}\sum_{i=1}^{\nu}\mathbb{1}(T(s_i)=t_0)$ to $\mathbb{P}(X=t_0)$ as used below equation (3).

Therefore, under H_0, by a property of almost sure convergence we have that

$$\begin{aligned}
\mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid \text{(ii)}\right) &= \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0 < t_{b-1},\ \text{(ii)}\right)\mathbb{P}\left(t_0 < t_{b-1}\right) + \\
&\quad \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0 > t_{b-1},\ \text{(ii)}\right)\mathbb{P}\left(t_0 > t_{b-1}\right) + \\
&\quad \mathbb{P}\left(\text{reject } H_0 \text{ on left tail} \mid t_0 = t_{b-1},\ \text{(ii)}\right)\mathbb{P}\left(t_0 = t_{b-1}\right) \\
&\xrightarrow{a.s.} 1 \cdot F(t_{b-1}) + 0 \cdot (1 - F(t_b)) + \rho \\
&= F(t_{b-1}) + \rho \\
&= F(t_{b-1}) + \alpha/2 - F(t_{b-1}) \\
&= \alpha/2.
\end{aligned} \tag{9}$$

The proof is symmetric for the probability of a rejection on the right tail of the empirical distribution of $T(s)$.

With H_0 asymptotically being rejected with probability $\alpha/2$ on both ends of the empirical distribution of $T(s)$ under the null, H_0 is rejected with probability α under the null given the rule to reject the null when $2\min\left\{\frac{|l|+|e_l|}{\nu}, \frac{|g|+|e_g|}{\nu}\right\} < \alpha$. ■

2.3. Multiple Testing

In appendix section A I show how the test generalizes to multiple test statistics computed per sample. The extremity of the values of multiple test statistics per selected sample on the population graph may namely be more indicative of sample representativeness of the population on average than the extremity of the value of just one of these test statistics.

3. Application

3.1. Social Network Study

School Experiment

I now turn to the application of the test with real-world data. The setting is a randomized experiment of student-specific anti-conflict training with 56 public middle schools is New Jersey (Paluck, Shepherd, and Aronow 2016). In each of the 28 treated schools, a strict subset of students are randomly selected to participate in the experiments. The researchers have all students in the 56 schools fill out a survey before and after the intervention. The surveys ask each student to list which other students they interact with. One can construct a social network with the survey responses.

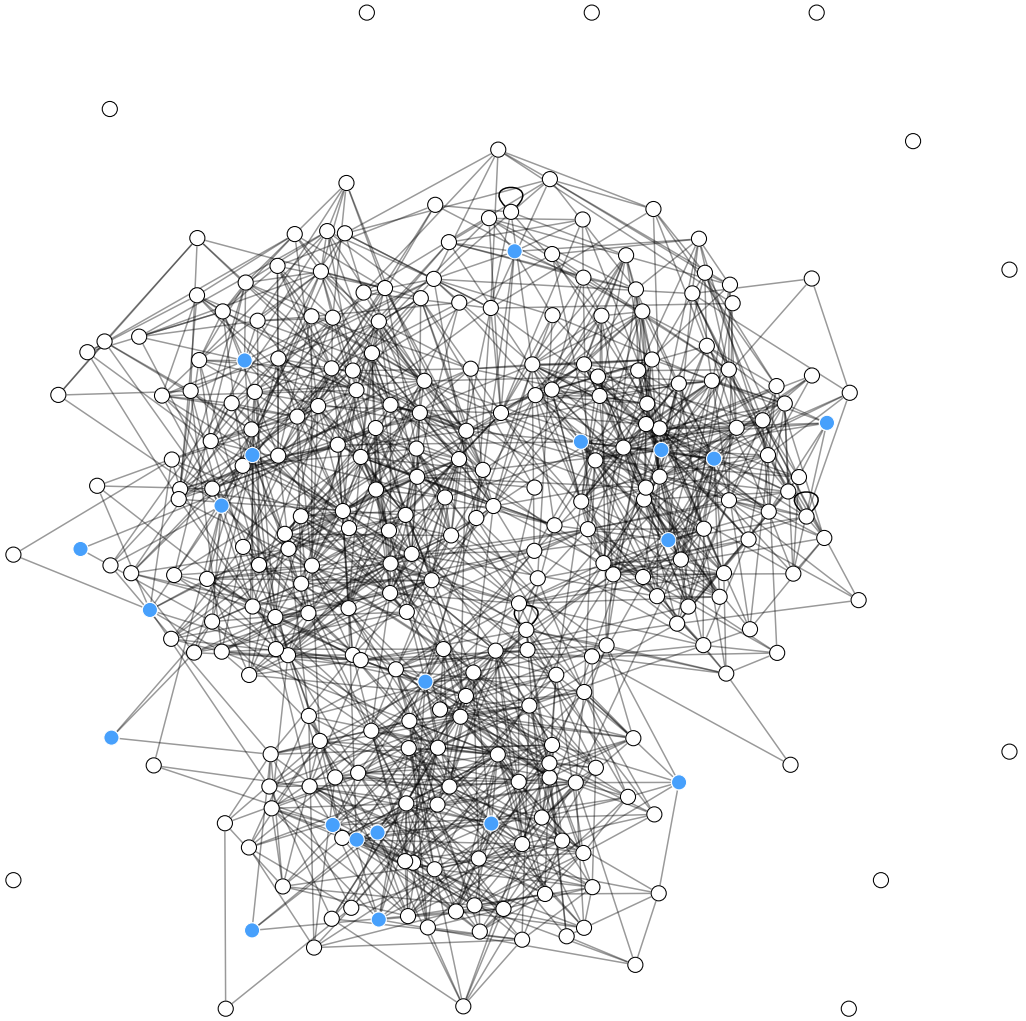
In this case one can use a t-test on demographic variables to test whether the treated sample is representative of the population sample (i.e., all students in a middle school). With that, a sample representativeness test on the network of the treated students may not add that much information about the representativeness of the selected sample. I do, however, apply the graph representativeness test on this data because of two properties. First, there are multiple populations (schools) among which treatment is randomly assigned[5]. If the researchers implemented the randomization correctly, then the representativeness test on the network of treated students can

[5]To be more precise, students were randomly selected to be eligible to be treated from a strict subset in each school, but could decide themselves in agreement with their parents whether to accept treatment.

be expected to be rejected for a proportion close to α of all schools. Second, the availability of population demographic data (i.e., student's demographic data for each school) allows comparing the results from a network representativeness test with the results from demographic representativeness test carried out on the same selected samples. Use cases will follow for which t-tests on demographic variables can not be carried out due to data availability issues.

See the below for an example of the social network defined on one school's student population.

Figure 2: Friend Network in Particular School



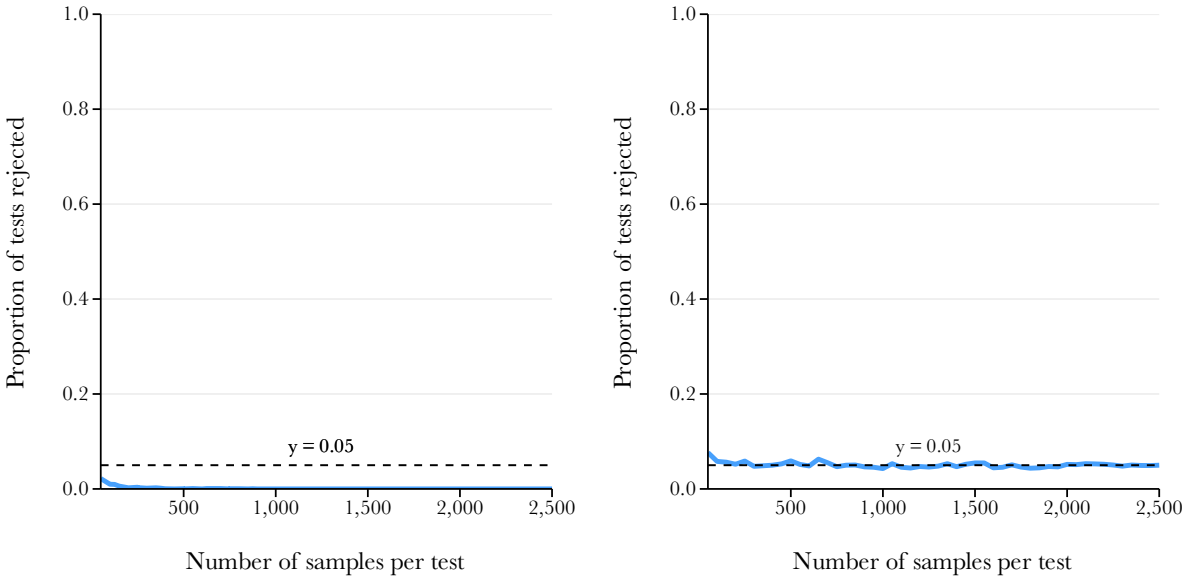
Notes: This figure presents an example of a school's friend network in the format of a colored undirected graph. The nodes are students. An edge between two students exists whenever one of them submitted the other as someone they spent time with in the last few weeks. I use responses only from the baseline survey. I color nodes in blue when the corresponding students receive anti-conflict training. You can observe three friend clusters in this graph, which represent all three grades of students in the given school. One can imagine that selection into the survey of students with similar interests to serve the school can give rise to more extreme graph statistic than randomly selected samples typically have, due to a higher degree of connectivity between students with similar interests than between randomly selected students.

For each sample of 50 individuals with the corresponding induced subgraph s_i, I define the test statistic $T(s_i)$ as the number of components in s_i.

The below two figures show the empirical probability of rejecting the null hypothesis at the level $\alpha = 0.05$ with the number of samples ν growing large as required here. The rejection

probability converges to 0 for the samples that selected into treatment, decreasing confidence that the randomization of treatment in the experiment was correctly implemented, but perhaps increasing the confidence that the students selected into treatment are representative of the population. Furthermore, the rejection probability for a random selection of survey participants seems to converge to 0.05 as expected.

Figure 3: Empirical Rejection Probabilities for Increasing ν: Social Network Study



(a) Selected Sample

(b) Random Sample

Notes: This figure presents the empirical rejection probabilities of H_0 at the 0.05 significance level for the two sample types as the number of samples ν grows large. I conduct 2,500 tests per value of ν, such that I re-select a reference sample for each test.

Related Use-Case

One can apply an approach akin to the above for a survey conducted with respondents selected from a social media platform. Suppose I want to estimate average attitudes towards some statements among the active US Facebook user population that follow the Republican Party page. I may then recruit eligible respondents through a Facebook advertisement or a survey vendor.

Normally, I may present a table of t-test for a collection of demographic variables to show whether there is evidence that the survey sample is significantly different from the population. I can collect demographic information on the survey sample in my survey. However, I may not be able to collect extensive reliable demographic information on the sampling population (in this case people who follow the Republican Party on Facebook).

A lot of the Facebook network structure[6] is publicly observable, though. I can then compare a statistic computed on my survey sample with that computed on a collection of random samples from the sampling population.

[6] I define the "Facebook network structure" as vertices being profiles, and directed edges indicating whether one profile follows another. There are other ways to define the network structure.

3.2. Uber Internal Rider Survey

Overview

A company like Uber may be interested in the experience of the average riders regarding the platform in some large city, like New York City. They may invite riders in that area to complete a survey. The demographic information on out-of survey sample Uber riders probably varies, and may at worst just be billing information from a friend of the rider plus a name (perhaps a pseudonym!).

In this case, Uber can refer to the trip history of riders in the survey sample and population to say something about the representativeness of the sample. If a statistic of the riding behavior of your selected sample is extreme in a collection of the same statistic for randomly selected samples of riders, your survey sample can be expected to feature an excess of non-standard individuals. For each rider, it can determine how many times each road segment in NYC has been traversed. This defines a weighted graph, where nodes are road segment endpoints, edges are road segments, and weights are traversal counts per road segment. Uber can then sum the n weighted graphs of the selected sample and define a test statistic on such an aggregated version of a weighted graph.

Besides t-tests on simple rider information like the number of trips taken and average trip length, looking at a representativeness test on the graph structure of rides can provide richer evidence that the sample is not representative.

Results

To put the above into practice, I start by simulating the riding history of 237 Uber riders living each in one of 237 Census block groups in lower-Manhattan. Each Census block group is home to one rider in this simulation. The number of trips a rider takes is proportional to the per capita income in the Census block group that they are from because richer individuals will be able to afford an Uber more often. I simulate trips of type (a), (b), and (c). Trips (a) start from the rider's home Census block group and end in another with 35% probability, (b) end in the rider's home Census block group and start in another with 35% probability, and (c) start and end in distinct non-home Census block groups with remainder probability. The probability of starting or ending in some non-home block group is negatively proportional to its centroid's distance with the home Census block group centroid and is further lowered when the block group is too close to home[7].

Furthermore, the probability of selecting into the "Uber" rider survey is negatively proportional to the square root of per capita income in the rider's home Census block group. I motivate this with the idea that riders with a lower income are more likely to be incentivized to complete a survey for a fixed cash/coupon/discount reward.

[7] You probably will be less likely to take an Uber from/to places close enough to home.

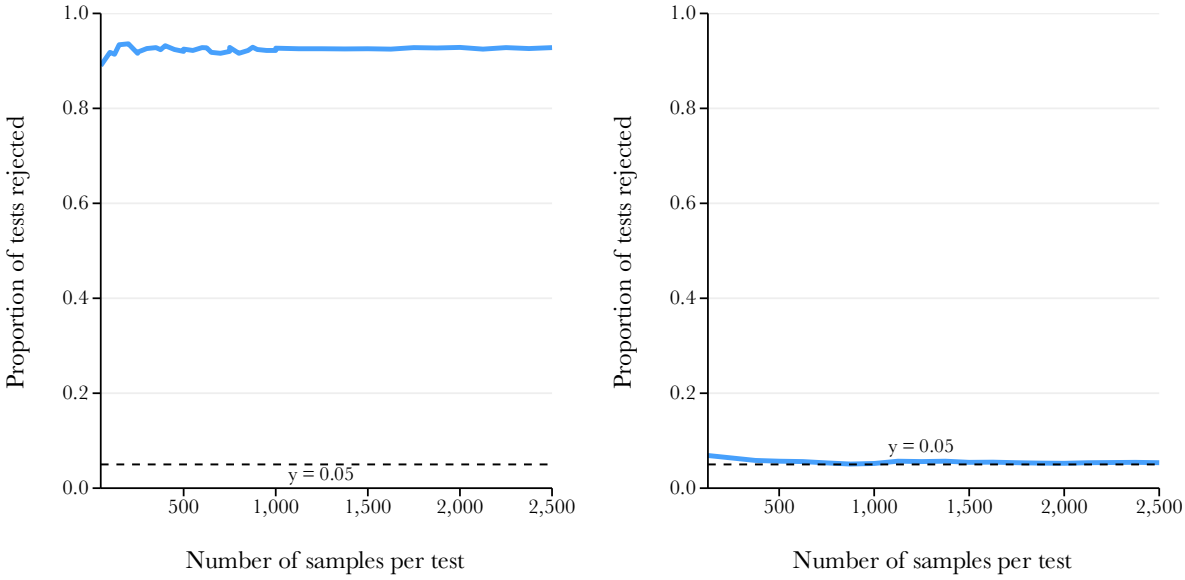
Figure 4: Trips per Road Segment for Selected Sample

Notes: This figure presents the number of trips per road segment in lower-Manhattan for an example sample of 50 riders who selected into the survey.

For each sample of 50 individuals that defines the aggregated riding history s_i, I define the test statistic $T(s_i)$ as the average of distance to the center of lower-Manhattan during the Uber rides. Samples of people who ride their Ubers around closer to the lower-Manhattan centroid may have a non-typical job or interest more than riders from a randomly selected sample would have, on average. See section B in the appendix for the derivation of the test statistic.

The below two figures show the empirical probability of rejecting the null hypothesis at the level $\alpha = 0.05$ with the number of samples ν growing large as required here. The rejection probability seems to converges to 1 for samples that select into the survey. This is desirable as we know that the income distribution of these riders is much different from the population income distribution by construction. At the same time, the rejection probability for a random selection of survey participants seems to converge to 0.05.

Figure 5: Empirical Rejection Probabilities for Increasing ν: Uber Rider Survey



(a) Selected Sample

(b) Random Sample

Notes: This figure shows the empirical rejection probabilities of H_0 at the 0.05 significance level for the two sample types as the number of samples ν grows large. I conduct 2,500 tests per value of ν, such that I re-select a reference sample for each test.

References

Paluck, Elizabeth Levy, Hana Shepherd, and Peter M. Aronow (2016). "Changing climates of conflict: A social network experiment in 56 schools". In: *Proceedings of the National Academy of Sciences* 113.3, pp. 566–571. DOI: 10.1073/pnas.1514483113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1514483113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1514483113.

Ritzwoller, David M., Joseph P. Romano, and Azeem M. Shaikh (2024). *Randomization Inference: Theory and Applications*. arXiv: 2406.09521 [econ.EM]. URL: https://arxiv.org/abs/2406.09521.

Appendix

Table of Contents

A. Multiple Testing

In some situations, the extremity of the values of $\kappa > 1$ different test statistics per selected sample on the population graph is more indicative of sample representativeness of the population on average than the extremity of the value of just one of these test statistics. In that case, one can test the same hypothesis that the sample is randomly sampled from the population by following Algorithm 1, but adjusting the definition of X to now be the random variable corresponding to the index of the convex hull the values of the test statistics belong to that are computed for a random selection of n units. We also adjust the definition of l in Theorem 1 to become $l = \{i \in [\nu] : \eta(T(s_i)) < \eta(T(s_0))\}$, where η is the mapping returning the convex hull index a test statistic combination belongs to with $\eta : \mathcal{T} \mapsto \mathcal{H}$, $\mathcal{T} \subset \mathbb{R}^\kappa$, and $\mathcal{H} \subset \mathbb{N}$. We adjust the definition of the sets g and e in the same manner. Given the definition of X and η, $F(\tau)$ is now given by $F(\tau) = \mathbb{P}(X \leq \tau) = \sum_{t \in \mathcal{T}} \mathbb{1}(\eta(t) < \eta(\tau)) \mathbb{P}(X = t)$. $\hat{\mathbb{P}}$, $\mathbb{P}$, and $\hat{F}$ are adjusted similarly.

With these substitutions, the proof that we reject the null hypothesis with probability α under the null when using the rejection condition of $2\min\left\{\frac{|l|+|e_l|}{\nu}, \frac{|g|+|e_g|}{\nu}\right\} < \alpha$ is then identical to the proof of Theorem 1 in section 2, except that we need to replace the proof that $\hat{F}(\tau) \xrightarrow{a.s.} F(\tau)$. This follows below.

Lemma 3. *$\hat{F}(\tau) \xrightarrow{a.s.} F(\tau)$ as $\nu \to \infty$ under the new specification.*

Proof. We can first rearrange the equation for the empirical CDF.

$$\begin{aligned}
\hat{F}(\tau) &= \sum_{t \in \tilde{\mathcal{T}}} \mathbb{1}(\eta(t) < \eta(\tau)) \hat{\mathbb{P}}(X = \eta(t)) \\
&= \sum_{t \in \tilde{\mathcal{T}}} \mathbb{1}(\eta(t) < \eta(\tau)) \hat{\mathbb{P}}(X = \eta(t)) + \underbrace{\sum_{t \in \mathcal{T} \setminus \tilde{\mathcal{T}}} \mathbb{1}(\eta(t) < \eta(\tau)) \hat{\mathbb{P}}(X = \eta(t))}_{=0} \\
&= \sum_{t \in \mathcal{T}} \mathbb{1}(\eta(t) < \eta(\tau)) \hat{\mathbb{P}}(X = \eta(t))
\end{aligned} \tag{10}$$

Just like in the proof of Lemma 2, we note that by a similar argument of Lemma 1, a property of almost sure convergence, and the fact that $|\mathcal{T}|$ is finite and fixed, we obtain that $\hat{F}(\tau) = \sum_{t \in \mathcal{T}} \mathbb{1}(\eta(t) < \eta(\tau)) \hat{\mathbb{P}}(X = \eta(t)) \xrightarrow{a.s.} \sum_{t \in \mathcal{T}} \mathbb{1}(\eta(t) < \eta(\tau)) \mathbb{P}(X = \eta(t)) = F(\tau)$. ∎

To determine η after drawing ν random samples of n units, we start by composing all resulting unique values of the test statistic in a κ-dimensional point cloud. Then, we define the convex

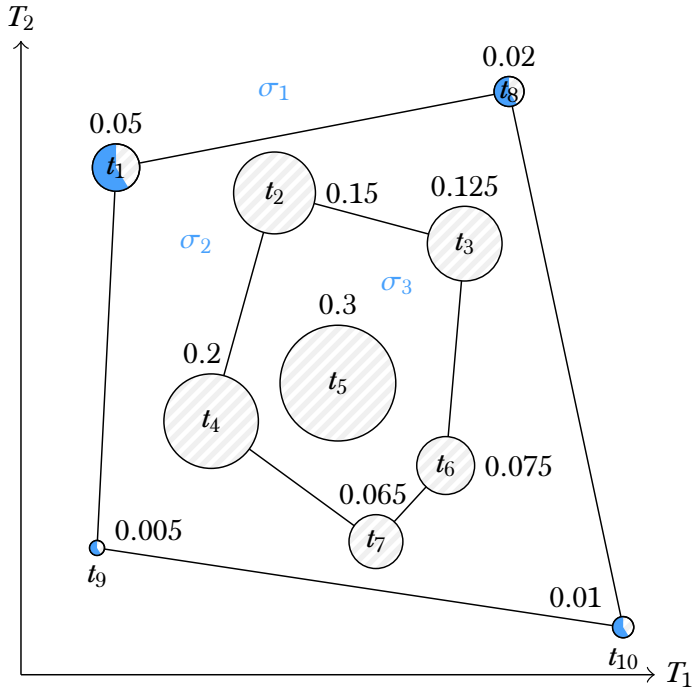
hull with index one to be the outer convex hull on this cloud of unique values. Hereafter, all values lying on that convex hull are removed, and the second convex hull is determined as the outer convex hull of the remaining unique values of the test statistic. This process repeats until there are no more points left. At that stage, we have determined for each test statistic t in $\tilde{\mathcal{T}}$ the convex hull index. The convex hull indices are collected in $\mathcal{H}$, a set counting up from one to the number of convex hulls drawn. See Figure 6 for a graphical illustration of the approach taken, for a dimensionality of $\kappa = 2$.

Definition 1 (Space enclosed by a convex hull). *The convex hull with index j, composed of the points in the set $\mathcal{T}_j = \{t_1, t_2, \ldots, t_\varsigma\}$, is given by* $\operatorname{conv}(\mathcal{T}_j) = \left\{\sum_{i=1}^{\varsigma} \omega_i t_i \;\middle|\; t_i \in \mathcal{T}_j,\ \omega_i \geq 0,\ \sum_{i=1}^{\varsigma} \omega_i = 1,\ \varsigma \in \mathbb{N}\right\}$. *With this, I define "the space enclosed" by this convex hull as*

$$\operatorname{conv}(\mathcal{T}_j) \setminus \underbrace{\left\{t \in \operatorname{conv}(\mathcal{T}_j) \;\middle|\; \forall \varepsilon > 0,\ \exists t' \in \mathbb{R}^\kappa \setminus \operatorname{conv}(\mathcal{T}_j) \text{ with } \|t' - t\|_2 < \varepsilon\right\}}_{\text{all points on the boundary of the convex hull}}.$$

I then define the space σ_i in the set $\{\sigma_1, \sigma_2, \ldots, \sigma_{|\mathcal{H}|}\}$ as the space that is enclosed by the convex hull with index $i-1$ minus the space enclosed by the convex hull with index $i \in \mathcal{H}$. If $i = 1$, σ_i is just the complement of the space enclosed by the convex hull with index i. Finally, we define $\eta(t) = i : t \in \sigma_i$. I choose to work with spaces instead of just hull boundaries because this guarantees the almost sure convergence of the empirical CDF to the CDF in Lemma 3. Taking a finite number of samples ν in practice does not guarantee we have encountered all unique combinations of test statistic values a sample of n units can take in the population. Namely, if we are conducting a test on the values of real-valued test statistics for samples of 50 social media users as part of a network of 1 million people, there are $3.3 \cdot 10^{235}$ unique samples you can draw, which may feature many unique combinations of values of the test statistics.

Figure 6: Graphical Illustration of Multiple Testing with $\kappa = 2$



Notes: This figure shows the unique combinations of values of the test statistics T_1 and T_2 that samples of size n in an imaginary population can take, displayed in separate circles. The size of the circles corresponds to the asymptotic probabilities a randomly drawn sample corresponds to the specific combination of values for all t_k in the population, part of the set $\mathcal{T} = \{t_1, t_2, \ldots, t_{10}\}$. The asymptotic probabilities add up to one. If the test statistic of the selected sample t_0 takes a value associated with a fully non-blue circle, the test will asymptotically not be rejected that the sample is randomly selected from the population. If t_0 takes a value associated with a circle with some blue in it, the test will asymptotically be rejected with the probability equal to the proportional coverage of blue in the corresponding circle. If the null hypothesis holds, we will asymptotically reject it with a probability equal to the proportion of space on the circles that is colored blue. In the picture, that proportion is 0.05. For this picture, we assume we know all test statistics in $\mathcal{T}$. I therefore do not visualize the probability the hypothesis is rejected when the selected sample's combination of test statistic values is not on the boundary of any of the convex hulls.

B. Rider Survey Test Statistic Derivations

B.1. Setup

I start by loading the geographical coordinates of road segment endpoints and project them to a local coordinate reference system. This will make it so that calculating distances between points at this stage will assume points lie on a flat plane. Thus, the calculated distances between two points in lower-Manhattan are an approximation of the true geodescic lengths (though a pedant might object that geodescic lengths would be an approximation too given the terrain geometries on which roads lie). In this case, we are calculating distances between points on a fairly small bounding box on earth's surface, lower-Manhattan, which means the approximations should be accurate enough for our liking.

I assume the points on a road segment lie on a straight line between the Cartesian coordinates of the road segment endpoints. With that, one can specify the Euclidian distance function of a

point across the road segment $r(\lambda)$ to a reference point c with:

$$d(r(\lambda),c) = \sqrt{(x(r(\lambda)) - x(c))^2 + (y(r(\lambda)) - y(c))^2}, \tag{11}$$

where $\lambda \in [0,1]$, $r(\lambda)$ returns the point on the road segment after having traversed proportion λ of it. The average distance of r to c will not depend on what endpoint of the road segment r corresponds to $r(\lambda = 0)$. The functions x and y take in a point and spit out the x-axis and y-axis coordinate, respectively.

I start with calculating the average distance between a road segment r and the reference point c because the overall riding history's average distance to the reference point c is a weighted average over $\frac{\int_0^1 d(r(\lambda),c)\mathrm{d}\lambda}{1-0}$ of all r. This integral of $d(r(\lambda),c)$ from $\lambda = 0$ to $\lambda = 1$ is the average distance to c when travelling across r with constant velocity.

We can rewrite the distance function as:

$$\begin{aligned}
d(r(\lambda),c) =& \sqrt{(x(r(\lambda)) - x(c))^2 + (y(r(\lambda)) - y(c))^2} \\
=& \sqrt{\left(\lambda \underbrace{[x(r(1)) - x(r(0))]}_{R_x} + \underbrace{x(r(0)) - x(c)}_{C_x}\right)^2 + \left(\lambda \underbrace{[y(r(1)) - y(r(0))]}_{R_y} + \underbrace{y(r(0)) - y(c)}_{C_y}\right)^2} \\
=& \sqrt{(\lambda R_x + C_x)^2 + (\lambda R_y + C_y)^2} \\
=& \sqrt{R_x^2\lambda^2 + 2R_xC_x\lambda + C_x^2 + R_y^2\lambda^2 + 2R_yC_y\lambda + C_y^2} \\
=& \sqrt{\underbrace{\left(R_x^2 + R_y^2\right)}_{\alpha}\lambda^2 + \underbrace{(2R_xC_x + 2R_yC_y)}_{\beta}\lambda + \underbrace{C_x^2 + C_y^2}_{\gamma}} \\
=& \sqrt{\alpha\lambda^2 + \beta\lambda + \gamma}
\end{aligned} \tag{12}$$

The integral of $d(r(\lambda),c)$ with respect to λ is then given for:

$$\int d(r(\lambda),c)\mathrm{d}\lambda = \frac{(2\alpha\lambda + \beta)\sqrt{\lambda(\alpha\lambda + \beta) + \gamma}}{4\alpha} - \frac{(\beta^2 - 4\alpha\gamma)\log\left(2\sqrt{\alpha}\sqrt{\lambda(\alpha\lambda + \beta) + \gamma} + 2\alpha\lambda + \beta\right)}{8\alpha^{3/2}}, \tag{13}$$

where $\log(x)$ is a complex-valued logarithm.

B.2. Average distance statistic

With the average distances per road segment, I start weighting by time spent on each road segment. The time spent on a road segment is the length of the road segment divided by the average speed on the road segment multiplied by the number of trips taken through the road segment. I can calculate this quantity for each road segment by taking the road segment speed limit to be the average speed on the road segment. I finally take the weighted average of all road segment average distances to get the average distance μ to the reference point c given an entire riding history H over the road network $\mathcal{R}$:

$$\mu(H,\mathcal{R},c) = \frac{\sum_{r\in\mathcal{R}} w(r) \int_0^1 d(r(\lambda),c)\mathrm{d}\lambda}{\sum_{r\in\mathcal{R}} w(r)}, \tag{14}$$

where $w(r)$ is the weight of time spent on road segment r with $w(r) = h(r,H)\frac{l(r)}{v(r)}$, $h(r,H)$ being the number of times that the road segment is traversed in the riding history H, $v(r)$ the speed limit on road segment r, and $l(r)$ the length of the road segment r.

B.3. Location concentration statistic

For each rider traversing a road segment r at constant velocity, the gravitational center of the location is $\left(x\left(r\left(\frac{1}{2}\right)\right), y\left(r\left(\frac{1}{2}\right)\right)\right)$.

I define the location concentration of all trips by (1) computing the gravitational center of all trips, and (2) computing the variance of the distance at any randomly drawn rider location to the computed gravitational center. The location concentration of all trips is the result from (2).

I compute (1) with

$$c_g(H,\mathcal{R}) = \left(\underbrace{\frac{\sum_{r\in\mathcal{R}} w(r) x\left(r\left(\frac{1}{2}\right)\right)}{\sum_{r\in\mathcal{R}} w(r)}}_{\text{x-coordinate}}, \underbrace{\frac{\sum_{r\in\mathcal{R}} w(r) y\left(r\left(\frac{1}{2}\right)\right)}{\sum_{r\in\mathcal{R}} w(r)}}_{\text{y-coordinate}} \right), \tag{15}$$

which can be interpreted as an average of the road segment-specific gravitional centers, weighted by the total time spent on each time segment: $\underbrace{h(r,H)}_{\text{Number of trips on } r} \times \underbrace{\frac{l(r)}{v(r)}}_{\text{Traversal time of } r}$.

Quantity (2) is defined as

$$
\begin{aligned}
\mathrm{Var}\left(d\left(X,c_g\right)\right) &= \mathbb{E}\left[\left(d(X,c_g)-\mathbb{E}\left[d(X,c_g)\right]\right)^2\right] \\
&= \frac{1}{\sum_{r\in\mathcal{R}} w(r)} \sum_{r\in\mathcal{R}} w(r) \int_0^1 \left(d(r(\lambda),c_g)-\mathbb{E}\left[d\left(X,c_g\right)\right]\right)^2 \mathrm{d}\lambda \\
&= \frac{1}{\sum_{r\in\mathcal{R}} w(r)} \sum_{r\in\mathcal{R}} w(r) \int_0^1 \left(d(r(\lambda),c_g)-\frac{1}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r') \int_0^1 d(r'(\lambda'),c_g)\mathrm{d}\lambda'\right)^2 \mathrm{d}\lambda \\
&= \frac{1}{\sum_{r\in\mathcal{R}} w(r)} \sum_{r\in\mathcal{R}} w(r) \int_0^1 \left(d(r(\lambda),c_g)-\frac{1}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r')\delta(r',c_g)\right)^2 \mathrm{d}\lambda \\
&= \frac{1}{\sum_{r\in\mathcal{R}} w(r)} \sum_{r\in\mathcal{R}} w(r) \int_0^1 \left(d(r(\lambda),c_g)^2-\frac{2d(r(\lambda),c_g)}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r')\delta(r',c_g)+\right. \\
&\quad \left.\left(\frac{1}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r')\delta(r',c_g)\right)^2 \mathrm{d}\lambda\right) \\
&= \frac{1}{\sum_{r\in\mathcal{R}} w(r)} \sum_{r\in\mathcal{R}} w(r) \left(\Delta(r(\lambda),c_g)-\frac{2\delta(r,c_g)}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r')\delta(r',c_g)+\right. \\
&\quad \left.\left(\frac{1}{\sum_{r'\in\mathcal{R}} w(r')} \sum_{r'\in\mathcal{R}} w(r')\delta(r',c_g)\right)^2\right),
\end{aligned}
\tag{16}
$$

where X is a randomly drawn location on the road network $\mathcal{R}$ of the riding history H, $\Delta(r(\lambda),c_g) = \int_0^1 \alpha\lambda^2+\beta\lambda+\gamma \mathrm{d}\lambda = \frac{1}{3}\alpha+\frac{1}{2}\beta+\gamma$ using the notation of equation (12), and $\delta(r,c_g)$ being the solution to $\int_0^1 d(r(\lambda'),c_g)\mathrm{d}\lambda$ from equation (13).