Investigating Transport for London Shared Bike Scheme: A Statistical Analysis of Demand Distributions During Peak Commute Hours

Group 18 Student IDs:
21140985
22221723
22061019
22149416
22197823

December 6, 2022

1 Introduction

In a world shaped by the impending doom of climate change, lukewarm post-pandemic outlooks and unavoidable plunge into economic crises, looking for alternatives to the habitual ways of living comes as an expected response. Bike sharing systems evolved as an alternative mode of transportation over half a century ago - but it is only in the recent decades this communal way of conveyance arose [11]. A 2021 Transport for London report estimates a 42% increase in using bicycles as the main mode of travel in "estimated average daily number of trips (millions)" from 2011 to 2020; a more recent span of 2019 to 2020 shows a yearly 6.7% change [5].

One benefit of cycling is its 'greenness' compared to vehicles. Nevertheless, bike sharing systems build up on their sustainable nature. Additional positive effects reported are less congestion, noise reduction, health benefits, and flexible mobility [11]. However, these assume consumers have *purposefully* switched from cars to shared bikes. Nevertheless, studies show no empirical evidence to support this assumed truth [7], leaving the advantage of bike sharing systems ambiguous although with great potential. Thus, this investigation aims to provide insights into the demand for bike sharing in London - and its popularity among commuters - in hopes of understanding the current state of the TfL's bike sharing scheme.

1.1 Problem Overview

As TfL's shared bike usage increases, studying the distribution of the bike rentals helps to uncover how their demand behaves. Specifically, this report focuses on the spring and summer peak evening commuting hours as a quick glance at the data shows that this is when the demand is at its highest.

The distribution of bike hires during the aforementioned period is studied first in Section 2, using goodness-of-fit tests to verify if the distribution is normal. Normal distribution has a crucial impact in statistics and is often used to represent real-valued random variables with unknown distributions [12]. It is also the theoretical basis for many statistical methods. Studying distributions can benefit by inferring probabilities of events, frequency distributions, and develop confidence intervals for statistical quantities. Consequently, it can help TfL understand an expected range of bike usage by applying statistical tests. In the following section, Q-Q plot first visually checks and then Anderson-Darling and Shapiro-Wilk tests formally examine the normality at significance level of 5%.

Section 3 inspects the differences between spring and summer distributions. Season can be an influencing factor for bicycle usage through variation in weather conditions, working, and commuting patterns. We examine spring and summer only and find it reasonable to assume the usage will be higher than in autumn and winter. Exploring if the distributions come from the same population provides a reference for TfL as to whether to respond to higher demand and allows for season-specific actions. Two-sample variants of Kolmogorov–Smirnov and Cramer-von Mises tests are used to study general differences in the distributions at significance level of 5%.

1.2 Exploratory Data Analysis and Assumptions

The dataset contains hourly count of rental bikes between years 2011 and 2012 in the Capital bikeshare system with the corresponding weather and seasonal information. For our analysis, we extract data of hours 16, 17 and 18 for spring and summer seasons, as these hours are defined as peak evening hours by TfL, also disregarding holidays and weekends. Thus we have m spring observations $X_1, X_2, \ldots, X_m$ where X_i is the number of bike hires in a peak commuting hour in spring. Similarly for summer we have n summer observations $Y_1, Y_2, \ldots, Y_n$. We assume, for each season, the number of bike hires in any peak evening hour on any commuting day is independent. So we have each X_i identically and independently distributed from distribution X, similarly Y_j is i.i.d from distribution Y. To summarise, we

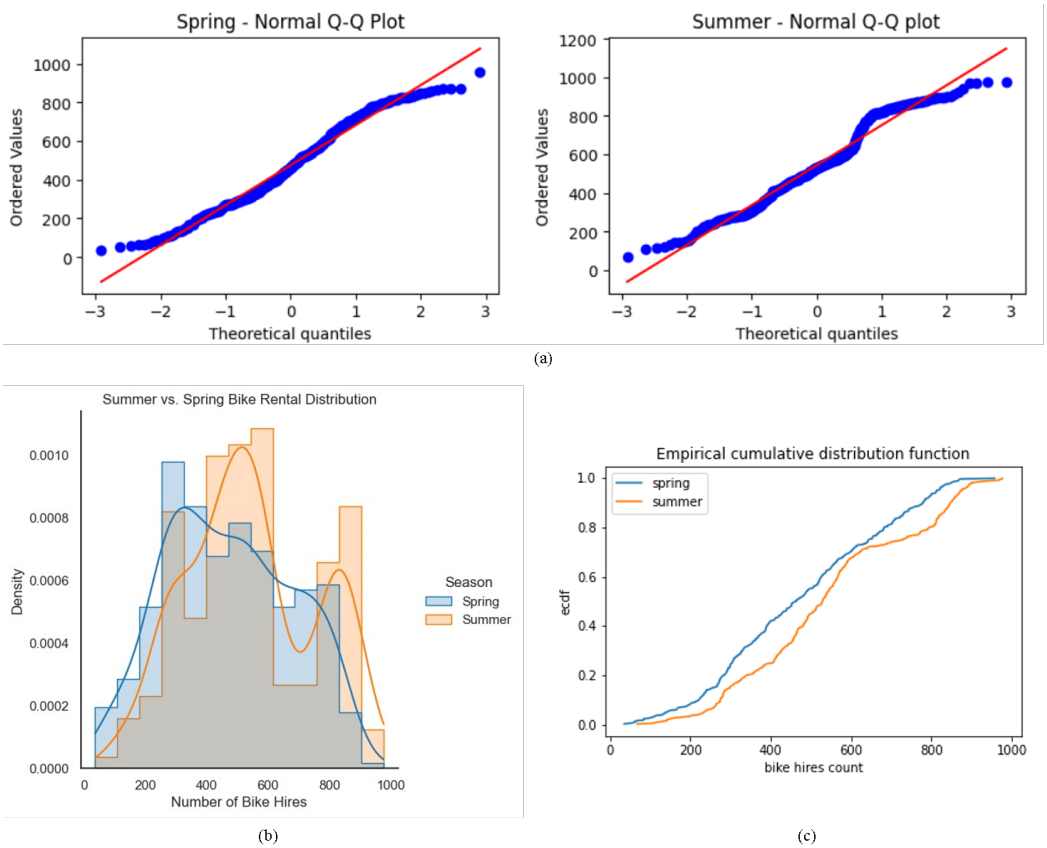

Figure 1: Data Exploration. (a) Normal Q-Q plot (b) Histogram (c) ECDF

have 384 and 393 observations for spring and summer respectively, with values between 36 and 957 for spring, and between 68 and 977 for summer.

To help us visually understand our data we plot normal QQ-plots and histograms, shown on fig. 1. The points on the QQ-plot for spring data can be seen to have less deviation from the normal line at the centre but more at the ends, indicating non-normal tails. For summer data, the points not only deviate from the straight line around the tails, but also around the centre which implies the data could be bimodal. It is reasonable to suspect that the data for both spring and summer is non-normal. Additionally, fig. 1(c) illustrates that the empirical cumulative distribution functions for the data of spring and summer are visually different.

2 Are bike hires during peak evening commuting hours normal?

We want to know if the number of bike hires in each evening peak commuting hour comes from a normal distribution. Stating this formally, our hypotheses for spring are:

$$H_0\text{: } X \text{ is normally distributed ; } H_1\text{: } X \text{ is not normally distributed}$$

We have the equivalent hypotheses for the summer distribution Y. According to Section 1.2 there was reason to suspect X and Y are non-normal. To complement this we use two Goodness of Fit tests to get a more quantified answer on our hypothesis.

Note, we are not assuming the shape or scale of the normal distribution. Both chosen tests are powerful across a range of different distributions [8].

2.1 Anderson-Darling Test

Our first test is the modified Anderson-Darling test which is a quadratic Empirical Cumulative Distribution Function (ECDF) test. It is comparing the squared distance between

the ECDF of our sample and the normal CDF using estimated mean and variance from our sample data. The test is defined as [6]:

$$A^* = n\left(1 + \frac{3}{4n} + \frac{9}{4n^2}\right)\int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dF(x) \tag{1}$$

where $F_n(x)$ is the ECDF of our sample data, and $F(x)$ is the CDF of $\mathcal{N}(\bar{X}, S^2)$ where $\bar{X} = \frac{1}{n}\sum_i X_i$ and $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$

A^* is a statistic of how far $F_n(x)$ is from $F(x)$. Because of the denominator of the integral the test has more weight at the tails. So $A*$ is larger when the tails of the distribution are non-normal. This weighting of the tails can cause the test to be less powerful if there are outliers in the sample. For an additional knowledge-expanding example on how to calculate A^* please refer to [10].

To estimate the critical regions and p-values, Monte Carlo methods have been used and extensive tables with the results can also be found in [10], and a derivation of the null distribution can be found in [6]. A significance level of 5% gives a critical region of $(0.779, \infty)$ for both summer and spring. We calculate A^* to be 2.52 and 4.44 for spring and summer respectively. Therefore we can **reject** H_0 for both both spring and summer, indicating that neither distributions were sampled from a normal distribution.

2.2 Shapiro-Wilk Test

Shapiro-Wilk test was chosen as a complement to the AD because SW focuses more centrally compared to AD. The test statistic W which takes values between 0 and 1 is calculated as follows[9]:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \times x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

where $(a_1, .., a_n) = \frac{m^T \times V^{-1}}{C}$, C is Vector Norm, m is vector of the order statistics i.i.d. random variable samples from standard normal distribution, V is covariance of the order statistics, $x_{(i)}$ is ith ordered sample.

Intuitively, the term inside the parenthesis in the numerator of W is interpreted as the slope of observed data versus the expected normal value, normalised to constant. It is the slope of QQ-plot squared and if the data is normally distributed then the numerator would be an estimate of variance (σ^2). The denominator is also an estimate of population variance which means $W = 1$ would indicate H_0 to not be rejected [13]. If $W < 1$, then the *p–value* (approximated using Monte-Carlo) would tell us the approximate probability of observing W as extreme or more extreme than observed. For **spring** evening hours, we obtain $W = 0.976$ and *p–value* = 6*e*–06 and for **summer** data $W = 0.966$ and *p–value* = 6*e*–08. In **both** cases, the *p–value* is less than 0.05 which means we **reject** H_0 for both summer and spring meaning there is strong evidence that the data for both seasons do not follow a normal distribution. One limitation of this test is that it is weak against symmetric moderately long-tailed (Leptokurtic) distributions [9], however we are testing for normal distribution which does not fall into this category of distributions.

3 Are the distributions for spring and summer the same?

Following the above results, this section investigates if the bike counts for spring and summer come from the same underlying distribution. No underlying distributions are assumed hence we apply two *non-parametric* tests that do not assume any prior knowledge of the distribution of the data. We get the following hypothesis test for whether they are the same:

H_0 : X is of the same distribution as Y ; H_1 : X is not of the same distribution as Y

The test statistic for Kolmogorov–Smirnov Test and Cramer-von Mises Test are based on emprical distribution functions. Let $F_X(x)$ and $F_Y(x)$ be the empirical distribution function for the bike count data during spring and summer. Fig. 1(c) illustrates the ECDFs of data during spring and summer, which show different patterns.

3.1 Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov (KS) test analyses whether two underlying probability distributions differ in the terms of all characteristics of a distribution including location, dispersion and shape. The test statistic D for KS test is simply the maximum vertical (Euclidean) distance between the two ECDFs [1, 4],

$$D = \sup_x |F_X(x) - F_Y(x)| \tag{3}$$

Based on a two-tailed test, testing of the null hypothesis proceeds by comparison of $p\text{–}value$ against the level of test. We obtain $D = 0.1725$ and $p\text{–}value = 1.5533e - 5$. Because $p\text{–}value < \frac{1}{2}(1-\alpha) = 0.025$, we **reject** H_0 that the underlying distribution of bike usage during spring and summer is the same at the 95% confidence level. A limitation of the KS test is that it is more sensitive to deviations near the center of the distribution [4].

3.2 Cramer-von Mises Test

The second two-sample test is the Cramer-von Mises test. Similar to the AD test, CvM is also based on squared difference between the two ECDFs. The test statistic is given by [2]:

$$T = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_X(x) - F_Y(x)]^2 dF_{X+Y}(x) \tag{4}$$

Where n and m are respectively the spring and summer sample sizes and $F_{X+Y}(x)$ is the ECDF of the two samples combined. As seen from the statistic, CvM compares the two empirical distribution functions by analyzing the sum of the squared differences between the two over the entire ECDFs i.e. at each point in the joint sample. This whole joint sample approach improves on the KS test that simply compares the singular maximum distance between the ECDFs. Turning to our analysis, we obtain $T = 7.5945$ and $p\text{–}value = 0.000500$. We are testing against a two-sided alternative hence our $p\text{–}value < \frac{1}{2}(1-\alpha) = 0.025$. This yields the same result as the KS: we **reject** H_0 at a 95% confidence level indicating that spring and summer do not come from the same underlying distribution.

4 Conclusion

First, all tests are performed under the assumption that counts are i.i.d. However, one can argue that counts are conditionally dependent (sequentially chained events). Second, we assumed a continuous distribution for counts, but in nature it is discrete. For the KS-and AD test statistics, the significance level is affected by ties within the groups for two-sample tests, which usually does not happen for continuous samples. We observed a few similar values but few relative to the sample size, however, it poses a weakness to the $p\text{–}value$ carried out in the test [3]. The CvM test handles this issue and is generally said to generate more powerful results than the KS test [3].

From our analysis, there is significant evidence that the distribution of spring and summer bike hires every hour during peak commuting times is not normal, and even further, their distributions are not the same. We suggest the distributions are non-normal because fig. 1(a) clearly shows this qualitatively, and in section 2, both of the high-power tests rejected Normality. Similarly, fig. 1(b) visually gives evidence that spring and summer are different distributions, and quantitatively both tests in section 3 reject that summer and spring are the same distribution with significance level 5%.

Even though our results give suggestions as to what the distributions are not, it is still useful knowledge. We recommend TfL complete additional analysis to understand the demand for bike hires and include more recent data. We would suggest investigating a wider range of distributions and seeing if they come from the same family of distributions for statistical analysis of events. And finally, a regression model utilising other suspected key covariates (e.g. weather metrics and location) could provide a deeper understanding of bike hire demand.

References

[1] Vance W. Berger and YanYan Zhou. *Kolmogorov–Smirnov Test: Overview*. Sept. 2014. DOI: 10.1002/9781118445112.stat06558. URL: https://doi.org/10.1002/9781118445112.stat06558.

[2] Donald A Darling. "The Kolmogorov-Smirnov, Cramer-von Mises tests". In: *The Annals of Mathematical Statistics* 28.4 (1957), pp. 823–838.

[3] Francesco Laio. "Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters". In: *Water Resources Research* 40.9 (Sept. 2004). DOI: 10.1029/2004wr003204. URL: https://doi.org/10.1029/2004wr003204.

[4] John Lanzante. "Testing For Differences Between Two Distributions In The Presence Of Serial Correlation Using the Kolmogorov–Smirnov and Kuiper's Tests". In: *International Journal of Climatology* 41 (May 2021). DOI: 10.1002/joc.7196.

[5] Transport for London. "Travel in London - Report 14". In: *Travel in London reports* (2021). URL: https://content.tfl.gov.uk/travel-in-london-report-14.pdf.

[6] Miodrag Lovric, ed. *International Encyclopedia of Statistical Science*. Springer, 2011. ISBN: 978-3-642-04897-5. DOI: 10.1007/978-3-642-04898-2. URL: https://doi.org/10.1007/978-3-642-04898-2.

[7] Peter Midgley. "Bicycle-sharing schemes: enhancing sustainable mobility in urban areas". In: *United Nations, Department of Economic and Social Affairs* 8 (2011), pp. 1–12.

[8] Xavier Romão, Raimundo Delgado, and Aníbal Costa. "An empirical power comparison of univariate goodness-of-fit tests for normality". In: *Journal of statistical computation and simulation* 80.5 (2010), pp. 545–591. ISSN: 0094-9655.

[9] Patrick Royston. "Approximating the Shapiro-Wilk W-test for non-normality". In: *Statistics and computing* 2.3 (1992), pp. 117–119.

[10] "Tests For The Normal Distribution". In: *Goodness-of-fit techniques / edited by Ralph B. D'Agostino, Michael A. Stephens*. Statistics, textbooks and monographs ; vol. 68. M. Dekker, 1986, pp. 372–374. ISBN: 0824774876.

[11] Mingshu Wang and Xiaolu Zhou. "Bike-sharing systems and congestion: Evidence from US cities". In: *Journal of Transport Geography* 65 (2017), pp. 147–154. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2017.10.022. URL: https://www.sciencedirect.com/science/article/pii/S0966692317302715.

[12] Eric W Weisstein. "Normal distribution". In: *Mathworld - A Wolfram Web Resource* (2002).

[13] Bee Wah Yap and Chiaw Hock Sim. "Comparisons of various types of normality tests". In: *Journal of Statistical Computation and Simulation* 81.12 (2011), pp. 2141–2155.

4.1 Plagiarism

We are fully aware of and adhere to the content of the "Plagiarism and Collusion" section in the Taught Postgraduate Student Handbook for the Department of Statistical Science.

4.2 Contributions

22061019 - Two sample test, introduction. **22149416** - Two sample test, conclusion. **21140985** - Introduction, goodness-of-fit test. **22221723** - Two sample tests, introduction. **22197823** - Goodness-of-fit test, conclusion.