5. Dec. 2022

Assessing the Distribution of Bike Hires with Goodness-of-Fit and Two-Sample Tests

Group ICA for STAT0032: Introduction to Statistical Data Science
UCL Department of Statistical Science

Group 19

Student Numbers: 22218945, 19002505, 22051936, 22092172, 22130975

1. Introduction

The UK is currently in the middle of an energy crisis, which has been building up over the last year as increased demand during the post-COVID economic reopening coincided with Russia-Ukraine war (Valero, 2022). Those events significantly impacted gas supplies and electricity and oil prices, further affecting the closely linked fossil-fuel-based transportation system (Greenpeace European Unit, 2022). According to a new published research (Greenpeace, 2022), alternative modes of transportation could reduce energy demand, thereby saving money and lowering emissions. This has resulted in renewed interest in bicycles in major cities such as London, which is partly supported by city-wide bike-sharing schemes. We are a team from a leading data science consultancy firm, and with this report, we provide relevant insights and recommendations for Transport for London's (TfL) bike-sharing scheme.

2. Problem Definition & Methodology

For the recommendations to TfL, we focus on the demand of bike hires during peak times and popular cycling seasons. Our analysis is based on the "Bike Sharing Data Set" (UCI, 2013) with records on bike hires in Washington D.C.. At the core of the analysis are two questions: (i) Does the number of bike hires during evening peak hours follow a normal distribution? (ii) Does the distribution of daily bike hires differ between spring and summer? We start with preprocessing the data set and explore first patterns. Next, we outline the procedure for each test used in the analysis. With three goodness-of-fit tests, we evaluate the first question, separately for spring and summer. To investigate the second question, we use two two-sample tests. Finally, we share the results from the statistical analysis and give recommendations for TfL.

3. Exploratory Data Analysis (EDA)

The dataset used in this analysis contains hourly and daily counts of bike hires between 2011 and 2012 within the bikeshare system of Washington D.C.. The original dataset consists of 17'379 hourly samples and 731 daily samples with no missing values. The target variables are the hourly and daily counts of bike hires and are both considered discrete with only integer values. This report looks into the hourly counts during peak times (4p.m. to 7p.m.) on working days only in spring and summer. Weekends and holidays are included in daily counts. It is worth noticing that 25% of the hourly samples in each season are duplicated which add up to a total of 181 duplicate values. The distributions of hourly counts in spring and summer are plotted in Figure 1. The distribution for spring roughly follows a bell shape while the distribution for summer has an additional peak on the right tail.

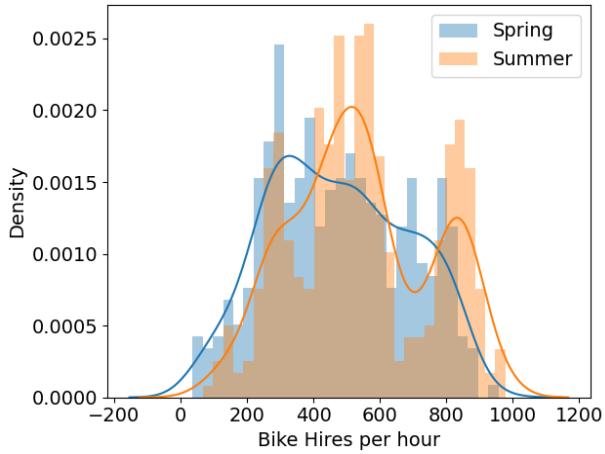

Figure 1: Distribution of hourly bike hires during peak hours in spring and summer

Figure 2 shows the distribution's deviation from normality at the tails for the hourly spring sample and we observed similar deviations for the summer sample.

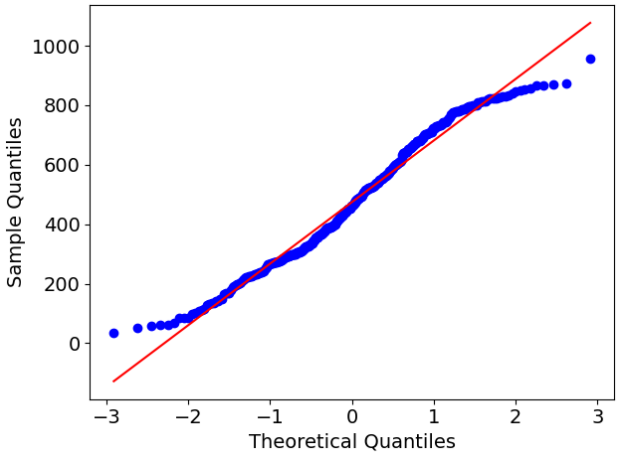

Figure 2: Q-Q plot comparing the quantiles of hourly bike counts in spring against those of a standard normal distribution

Distributions of the daily bike hire counts in spring and summer are shown in Figure 3. Similar to that observed in Figure 1, the distribution for summer also follows a double-peak shape due to a drop in frequency of counts at around 6'000. A much smaller drop is observed in the distribution for spring making the distribution very wide with a single peak. The mean daily bike hire counts in spring is around 4'992 which is smaller than that of around 5'644 in summer. However, the spring samples have a larger standard deviation of around 1'696 compared to that of 1'460 for summer.

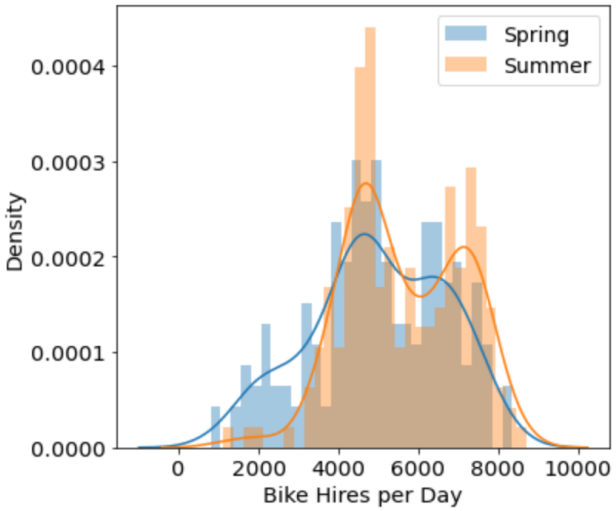

Figure 3: Distribution of daily bike hires in spring and summer

4. Goodness-of-Fit Tests

To assess the demand for bike hires during peak times, we use the following goodness-of-fit tests on both, the hourly bike sample in spring and in summer. These tests examine whether the sample approximately follows a particular distribution (Wasserman, 2004). We investigate for a normal distribution and the hypotheses are:

H_0 *: the sample follows a normal distribution*

H_1*: the sample does not follow a normal distribution*

4.1 Chi-Square Test

The Chi-Square (CSQ) is a non-parametric hypothesis test. Sample data is divided into intervals and the frequency is compared to the expected number of observations in each interval. Thus, CSQ is suitable for testing normality. The test determines how significantly the observed and expected frequency differ (Glen, 2017). The assumptions are: the variable to be tested is categorical and mutually exclusive; the expected number of observations per interval is ≥ 5; the intervals contain count cases and random samples (Statistics Laerd, n.d.). The test statistic is defined as:

$$CSQ = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{o_i}$$

O indicates the observed value and E is the expected value for each of the k intervals, and i is the 'i_{th}' interval. The distribution of the test statistics under the null is $CSQ \sim \chi^2_{k-1}$. However, the test statistic is dependent on how the data is binned and it is sensitive to the choice of bins. There is no optimal option for bin width. Also, the test is less powerful than other tests when the classification variable is continuous rather than discrete (Quality Control Plan, n.d.). It is still a useful and quick way of checking normality especially when we have discrete data points.

4.2 Anderson-Darling Test

The Anderson-Darling test (AD) compares the empirical cumulative distribution function (ECDF) of a sample with the normal cumulative distribution function (CDF). The test is sensitive in the tails of a distribution, and since our sample especially deviates in the tails, the test is suitable in assessing normality. The test statistic is calculated as:

$$AD = -N - \sum_{i}^{N} \frac{(2i-1)}{N} \left[lnF(Y_i) + ln(1 - F(Y_{N+1-i})) \right]$$

N denotes the sample size, F represents the CDF of the normal distribution and Y_i's are the observations in ascending order (Razali & Wah, 2011). The test assumes ordinal data and follows a numerically complex distribution under the null hypothesis (Marsaglia & Marsaglia, 2004). Thus, we used the adjusted test statistic and p-values according to Taeger & Kuhnt (2014), for which the distribution parameters are estimated. The test is not reliable if duplicate values exist (de Smith, 2018), which our sample has revealed. To address this, we added random uniform noise

$Unif(-1, 1)$ to the observations before calculating the test statistic.

4.3 D'Agostino-Pearson Test

The D'Agostino-Pearson (DP) test is used to assess the normality of data distribution using two statistical parameters, *skewness* and *kurtosis*. Skewness is the measure of asymmetry of data while kurtosis measures the peakedness in data. The test statistic is

$$DP = s^2 + k^2$$

where, s and k are zero-score, the ratio of skewness or kurtosis by their standard errors, of skew test and kurtosis test respectively (Kim, 2013). Under the null hypothesis, the skewness and kurtosis of data are zero, and the test statistics s^2 and k^2 follow standard normal distribution. This test is not the most powerful test for normality but with low kurtosis, symmetric and short-tailed data it is often recommended to use (Yap & Sim, 2011). This test for normality is preferred when dealing with large sample sizes along with repeated samples and our data has about 25% duplicate values for both summer and spring.

5. Two-Sample Tests

From the data, we observed that the daily bike hires in spring have a lower mean and larger variance than that in summer. In this section, two-sample tests were carried out to further investigate whether the distributions of daily bike hires in spring and summer are different. The null and alternative hypothesis are:

H_0: *the two distributions are the same*

H_1: *the two distributions are different*

The samples used in this section are assumed to be ordinal given that there are only 2 duplicate values.

5.1 Two-Sample Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (KS) test examines whether two sets of samples were drawn from the same distribution by comparing their ECDFs (Trevisan, 2022). The test assumes that all the samples are independently and identically distributed random variables and the two sample sets are mutually independent. The test statistic of the KS test is the maximum vertical distance D between the two ECDFs $F_1(x)$ and $F_2(x)$ of the samples:

$$D = max|F_1(x) - F_2(x)|$$

The test statistic D is sensitive to differences in both the location and shape of two ECDFs. It follows a zero-mean normal distribution under the null hypothesis. The critical value of the test statistic can be calculated using the significance level and the sizes of the sample sets. A major limitation of the KS test is that it is less sensitive to the differences in the tails of the distributions.

5.2 Mann-Whitney U Test

The Mann-Whitney (MW) test, also known as the Mann-Whitney-Wilcoxon test, is a non-parametric test. It is generally used to assess the differences between two independent groups when the dependent variable is ordinal or continuous but not normally distributed (Daniel, 1978). It follows the assumption that the data consist of a random sample of observations X_1, X_2, ..., X_{n_1} from population 1 with unknown median M_x, and another random sample of observations Y_1, Y_2, ..., Y_{n_2} from population 2 with unknown median M_y. The following formula computes the test statistic:

$$T = S - \frac{n_1(n_1+1)}{2}$$

S is the sum of the ranks assigned to the sample observations from population 1 and n_1 is the number of samples from population 1. The test statistic is said to follow an approximate normal distribution.

6. Results

The hypothesis test results are summarised in Table 1. The p-values are compared against the common significance level of $\alpha = 0.05$. It can be concluded that H_0 is rejected for all six goodness-of-fit tests as each p-value is less than $\alpha = 0.05$. Hence, we conclude that both spring and summer do not follow a normal distribution. These results support our observation from

Figures 1 and 2 where we witnessed indications that the hourly sample deviates from the normal distri- bution. As the number of bike hires is considered a discrete variable, it is also difficult to prove that the variable is continuously distributed.

Based on the results for the two two-sample tests, H_0 is rejected as well. At 95% confidence level, there is a significant difference between the distribution in spring and summer for daily bike hires.

Table 1: Tests for normality and for different distributions

Goodness-of-Fit Tests

Test	*CSQ*		*AD*		*DP*	
Season	*SP*	*SM*	*SP*	*SM*	*SP*	*SM*
Test Statistic	**48.87**	**154.92**	**2.52**	**4.45**	**51.12**	**39.32**
p-value	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*

Two-Sample Tests

Test	*KS*	*MW*
Test Statistic	**.18**	**13'672**
p-value	<.01*	<.01*

Significance level: * $p < 0.05$. Note: < .01 indicates $p < 0.01$, SP stands for spring, SM for summer. Each goodness-of-fit test was separately performed on the hourly bike sample in spring (384 observations) and in summer (393 observations). Two-sample tests were performed on daily bike hires (184 observations in spring, 188 observations in summer).

7. Conclusions

In this report, we conducted a statistical analysis on bike hires during peak hours and in different seasons to infer insights about the potential demand for TfL's bike sharing-scheme. The analysis was motivated by bike-sharing data in Washington D.C., and goodness-of-fit and two-sample tests were used to gain a distributional understanding of the data. We concluded that hourly bike hires are not normally distributed in spring and summer and that daily bike hires follow different distributions in those seasons.

For a plan of action, we recommend TfL to further investigate the distribution of the hourly samples. More precise knowledge about their distribution will be able to support in forecasting the bike demand at busy hours to prepare bikes supply. It will also be helpful for TfL if they investigate how exactly the two distributions differ. The EDA showed that summer might have a higher and more stable mean daily bike usage than spring. We advise TfL to adopt different promotion and supply strategies to adapt to the differences in demand for the two seasons. A method TfL can apply is by lowering the bike hire price in spring to boost demand.

In Figures 1 and 3, the hourly and daily bike samples demonstrate two peaks in summer. This could be indicated by cooperation with customers or competitors (Käki et. al., 2013) and we suggest conducting further research into that aspect. Besides further statistical analysis for better distributional knowledge, we also encourage incorporating expert knowledge in demand forecasting to optimise supply planning.

As the data used for this analysis are from 2011 and 2012 and based on bike hire in the USA., it is outdated and might not be suitable to refer to. It is suggested that TfL conduct further studies using more recent data or data from a city that is of similar size and population as London. It is also recommended that TfL conduct monthly maintenance on the bicycles so there will be enough supplies to meet the demand.

References

Daniel, W. W. (1978). *Applied Nonparametric Statistics*. Houghton Mifflin.

de Smith, M. J. (2018). *Statistical Analysis Handbook*. The Winchelsea Press.

Glen, S. (2017). Chi-square Test for Normality. *Statistics How To*. [online] Available at: https://www.statisticshowto.com/chi-square-test-normality/#google_vignette [Accessed 23 Nov. 2022].

Greenpeace. (2022). *Transport sector solutions*. [online] Available at: https://greenpeace.at/uploads/2022/09/transportsectorsolutions_report_by_greenpeace_cee_2022.pdf [Accessed 23 Nov. 2022].

Greenpeace European Unit. (2022). *How Europe's transport system can tackle the energy and climate crises*. [online] Available at: https://www.greenpeace.org/eu-unit/issues/climate-energy/46371/how-europes-transport-system-can-tackle-the-energy-and-climate-crises/ [Accessed 23 Nov. 2022].

Käki, A., Salo, A., & Talluri, S. (2013). Impact of the shape of demand distribution in decision models for operations management. *Computers in Industry*, ***64***(7), 765-775.

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, **38**(1), 52-54.

Marsaglia, G., & Marsaglia, J. (2004). Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, **9**, 1-5.

Quality Control Plan. (n.d.). *Does your data violate goodness of fit (chi-square) test assumptions?* [online] Available at: https://www.quality-control-plan.com/StatGuide/gf-dist_ass_viol.htm#Special%20problems%20with%20continuous%20classification%20variables [Accessed 23 Nov. 2022].

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, **2**(1), 21–33.

Statistics Laerd. (n.d.). *Chi-Square Goodness-of-Fit Test in SPSS Statistics*. [online] Available at: https://statistics.laerd.com/spss-tutorials/chi-square-goodness-of-fit-test-in-spss-statistics.php [Accessed 23 Nov. 2022].

Taeger, D., & Kuhnt, S. (2014). *Statistical Hypothesis Testing with SAS and R*. John Wiley & Sons, Incorporated.

Trevisan, V. (2022). Comparing sample distribution with the Kolmogorov-Smirnov (KS) test. *Towards Data Science.* [online] Available at: https://towardsdatascience.com/comparing-sample-distributions-with-the-kolmogorov-smirnov-ks-test-a2292ad6fee5 [Accessed 29 Oct. 2022].

UCI. (2013). Bike Sharing Data Set. *Centre for Machine Learning and Intelligent Systems*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset [Accessed 28 Oct. 2022].

Valero, A. (2022). Why have energy bills in the UK been rising?. *LSE British Politics and Policy*. [online] Available at: https://blogs.lse.ac.uk/politicsandpolicy/why-have-energy-bills-in-the-uk-been-rising-net-zero/ [Accessed 23 Nov. 2022].

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference / by Larry Wasserman*. (1st ed. 2004). Springer New York.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, **81**(12), 2141-2155.

Individual Contributions

22218945: section 2, 4, 6, 7, references
19002505: wrote part of section 3 and section 5
22051936: section 1 and 4
22092172: Section 3 and 4
22130975: Section 5, 6, 7

Plagiarism Statement

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