Name: Ayano Yamamoto

Dataset Used: Bike Sharing

Table of Contents

1. Does the temperature (actual) impact the total number of bikes hired per day?

State a hypothesis pair for the question

- Null hypothesis (H_0):
  Temperature (actual) does not impact the total number of bikes hired per day.
- Alternative hypothesis (H_1):
  Temperature (actual) impacts the total number of bikes hired per day.

Explore the data to be used to test the hypothesis

Outcome variable: `cnt`

`cnt` represents the count of total rental bikes including both casual and registered. It is a discrete quantitative variable with an interval scale since it can assume only a finite number of real values within the total number of bikes available for rental. The sample size is 731, and there is no missing data. The minimum is 22, the maximum is 8714, the mean is 4504, and the standard deviation is 1937.21.

`cnt` variable was assessed to see if it meets the requirements to be treated as a normal distribution. Inspection of the histogram with the density curve shows an approximate unimodal distribution with no strong skews, although the distribution is not smooth. The box plot shows a fairly symmetrical spread with no outliers (see Figure 1). The QQ plot shows an indication of thin tails on both sides (see Figure 2).

Figure 1: Distribution of total number of bikes hired per day

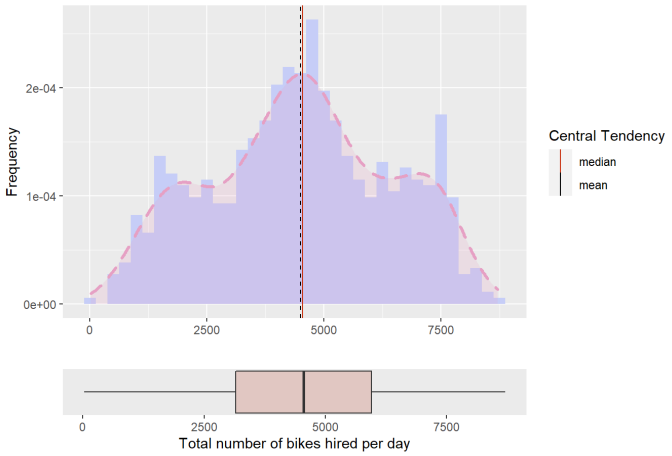
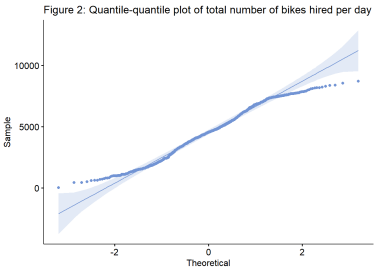


Figure 2: Quantile-quantile plot of total number of bikes hired per day

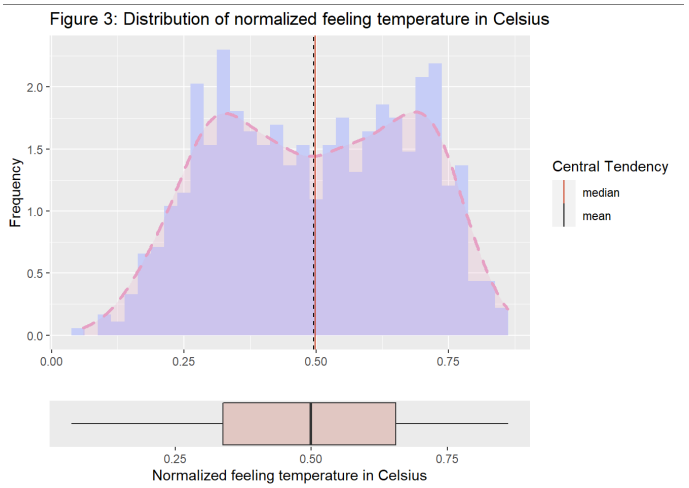


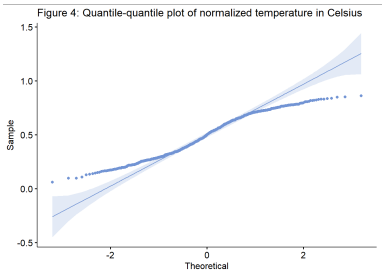
The quantified skewness of `cnt` is -.52 meaning the distribution is primarily symmetrical, and the quantified excess kurtosis is -4.48 which supports the platykurtic distribution seen in the QQ plot. Skewness is within ±2.00, but there is a potential issue with kurtosis. Examining the standardised scores shows 0% falling outside ±3.29, hence we will consider `cnt` an approximately normal distribution.

Independent variable: `temp`

`temp` represents the normalized temperature in Celsius. The values are derived via $(t - t\_min)/(t\_max - t\_min)$, $t\_min = -8$, $t\_max = +39$ (only on an hourly scale). It is a continuous quantitative variable since it represents a measurable amount in degrees Celsius. It is also an interval scale since the Celsius temperature scale has an arbitrary zero. The sample size is 731, and there is no missing data. The minimum is .06, the maximum is .86, the mean is .50, and the standard deviation is .18.

`temp` variable was assessed to see if it meets the requirements to be treated as a normal distribution. Inspection of the histogram with the density curve shows a bimodal distribution. From the box plot, the distribution seems mostly centred with a small left skew (see Figure 3). The QQ plot indicates that the distribution has thin tails on both sides with low outlier frequencies (see Figure 4).





The quantified skewness of `temp` is -.60 meaning the distribution is primarily symmetrical, and the quantified kurtosis is -6.17 indicating a platykurtic distribution. Skewness is within ±2.00, but there is a potential issue with kurtosis. Examining the standardised scores shows 0% falling outside ±3.29, hence we will consider `temp` an approximately normal distribution.

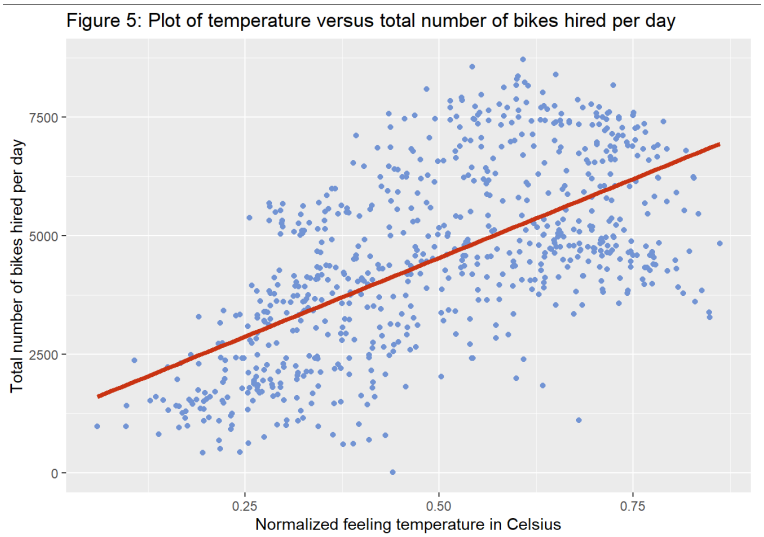
Identify the test to be used

We have a normally distributed outcome variable `cnt` and a normally distributed independent variable `temp`. We will use the Pearson correlation test to determine if the temperature (actual) impacts the total number of bikes hired per day.

Conduct the test and report your findings

From the scatter plot with a regression line, there appears to be a positive correlation between `temp` and `cnt` (see Figure 5).

Figure 5: Plot of temperature versus total number of bikes hired per day



For the p-value, we will use a common value of 5% (0.05). For effect size, Cohen suggested the conventional definitions of weak: r = ±.1, moderate: r = ±.3, and strong: r = ±.5 (Cohen p.83). Peck and Devore suggested differently, as weak: r < ±.5, moderate: r = ±.5, and strong r = ±8 (Peck and Devore p.216). In this project, we will use Cohen's effect size heuristic with observations on statistical significance and the amount of shared variance.

The relationship between the total number of bikes hired per day (count of total rental bikes including both casual and registered) and temperature (actual) (normalised temperature in Celsius) was investigated using a Pearson correlation. A statistically significant result was found indicating a strong positive correlation (r = .63, n = 73, p < .001). The coefficient of determination is $r^2 = .63^2 = .3969$ meaning the total number of bikes hired per day and temperature (actual) share 39.69% of their variance.

Based on these findings we will reject the null hypothesis (H_0) that the temperature (actual) does not impact the total number of bikes hired per day.

2. Does the level of humidity impact the total number of bikes hired per day?

State a hypothesis pair for the question

- Null hypothesis (H_0):
  The level of humidity does not impact the total number of bikes hired per day.
- Alternative hypothesis (H_1):
  The level of humidity impacts the total number of bikes hired per day.

Explore the data to be used to test the hypothesis

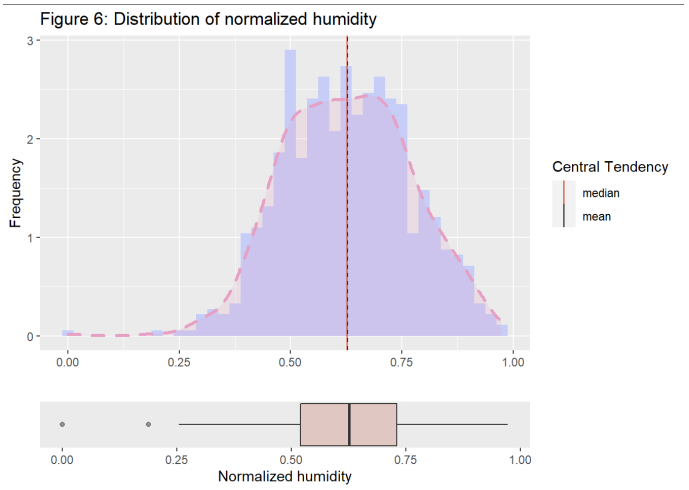
Outcome variable: `cnt`

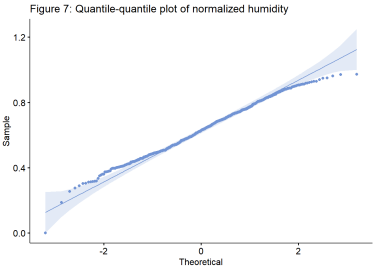
Please refer to the exploration from Question 1.

Independent variable: `hum`

`hum` represents the normalized humidity. The values are divided into 100 (max). It is a continuous quantitative variable since it represents a measurable amount of water vapour present in the air. It is also a ratio scale since there is a true zero in the scale when there is no water vapour present in the air. The sample size is 731, and there is no missing data. The minimum is 0, which is an unlikely number to be observed in nature. There is 1 observation (0.14% of sample size) with a recorded `hum` of 0. This observation is likely to be an error and could be considered for removal when cleaning the data for predictive models. The maximum is .97, the mean is .63, and the standard deviation is .14.

`hum` variable was assessed to see if it meets the requirements to be treated as a normal distribution. Inspection of the histogram with the density curve shows a unimodal distribution with outliers to the left. The box plot also shows that the distribution is skewed left (see Figure 6). The QQ plot indicates that the distribution has a fat tail on the left compared to the right, indicating a higher frequency of outliers on the lower end of the scale (see Figure 7).





The quantified skewness of `hum` is -.77, meaning it is approximately symmetric, and the quantified excess kurtosis is -.36 indicating an almost mesokurtic distribution. Since both skewness and kurtosis are within ±2.00, we will consider `hum` an approximately normal distribution. Inspection of the standardised score supports this, with .14% falling outside of ±3.29.

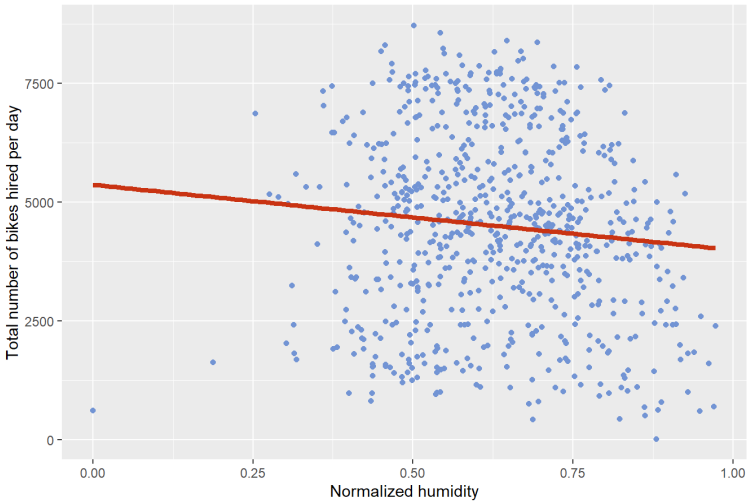
Identify the test to be used

We have a normally distributed outcome variable `cnt` and a normally distributed independent variable `hum`. We will use the Pearson correlation test to determine if the level of humidity impacts the total number of bikes hired per day.

Conduct the test and report your findings

From the scatter plot with a regression line, there appears to be a positive correlation between `hum` and `cnt` (see Figure 8).

Figure 8: Plot of normalized humidity versus total number of bikes hired per day



The relationship between the total number of bikes hired per day (count of total rental bikes including both casual and registered) and the level of humidity (normalised humidity) was investigated using a Pearson correlation. A statistically significant result was found indicating a weak negative correlation (r = -.10, n = 729, p = .01). The coefficient of determination $r^2 = -.10^2 = .01$ means the total number of bikes hired per day and the level of humidity share 1% of their variance.

Based on these findings we will reject the null hypothesis (H_0) that the level of humidity does not impact the total number of bikes hired per day.

3. Does the total number of bikes hired per day vary according to whether a day is a regular weekday or a weekend?

State a hypothesis pair for the question

- Null hypothesis (H_0):
  The total number of bikes hired per day does not vary according to whether a day is a regular weekday or a weekend.
- Alternative hypothesis (H_1):
  The total number of bikes hired per day varies according to whether a day is a regular weekday or a weekend.

Explore the data to be used to test the hypothesis

Outcome variable: `cnt`

Please refer to the exploration from Question 1.

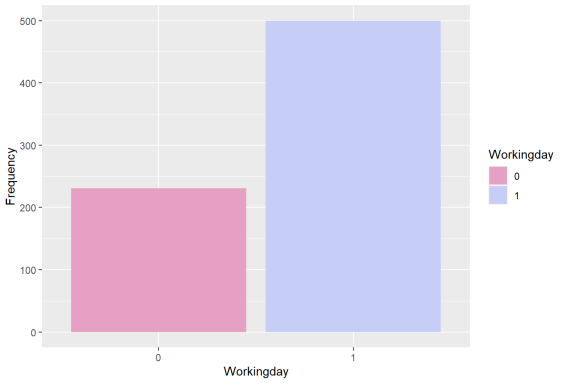
Independent variable: `workingday`

`workingday` represents if the day is a working day or not. If it is neither a weekend nor a holiday the recorded value is 1, otherwise it is 0. It is a nominal categorical variable since a day can only fit into one category. It is also binary (dichotomous) since the variable has only 2 groups, and those 2 groups are independent of each other. The sample size is 731, and there is no missing data. Possible values are 0 and 1. The most frequently occurring value is 1 which makes up 68.40% of the observations (see Table 1 and Figure 9).

Group	Frequency	Percentage
0	231	31.60%
1	500	68.40%

Table 1: Distribution of `workingday`

Figure 9: Distribution of whether a day is a regular weekday or a weekend



Identify the test to be used (justify your choice)

We have a normally distributed outcome variable `cnt` and a categorical variable of 2 independent groups `workingday`. Levene's test was conducted to investigate the homogeneity of variance. The null hypothesis is that the variances in groups are equal, so to assume homogeneity we would expect the probability to not be statistically significant. *Pr(>F)* is 0.04, which is under 0.05, meaning that variances are heterogeneous. We will use the parametric T test with `var.equal = FALSE` to determine if the total number of bikes hired per day varies according to whether a day is a regular weekday or a weekend.

Conduct the test and report your findings

An independent-samples t-test was conducted to compare the total number of bikes hired per day (count of total rental bikes including both casual and registered) and whether a day is a regular weekday or a weekend. No significant difference in the total number of bikes hired per day was found (M = 4330.17, SD = 2052.14 for weekends, M = 4584.82, SD = 1878.42 for regular weekdays), (t(413.94) = -1.60, p = .11). Cohen's d also indicated a very small effect size (-.16).

Based on these findings we will not reject the null hypothesis (H_0) that the total number of bikes hired per day does not vary according to whether a day is a regular weekday or a weekend.

4. Does the total number of bikes hired per day vary by the day of the week?

State a hypothesis pair for the question

- Null hypothesis (H_0):
  The total number of bikes hired per day does not vary by the day of the week.
- Alternative hypothesis (H_1):
  The total number of bikes hired per day varies by the day of the week.

Explore the data to be used to test the hypothesis

Outcome variable: `cnt`

Please refer to the exploration from Question 1.

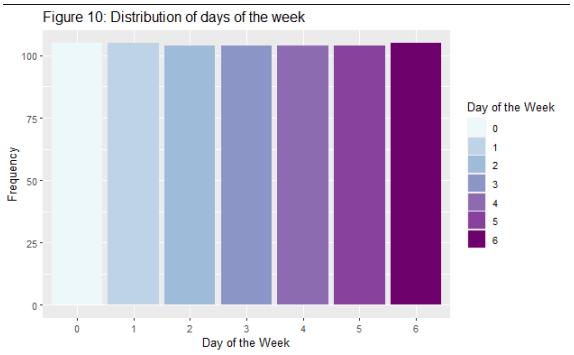
Independent variable: `weekday`

`weekday` represents the day of the week in integers. It is a nominal categorical variable since a day of the week cannot be ranked over or under another. Its groups are independent of each other. The sample size is 731, and there is no missing data. Possible values are 0, 1, 2, 3, 4, 5, and 6. The values are primarily equally distributed, with the most frequently occurring values 0, 1, and 6 making up 14.36% of the observations each (see Table 2 and Figure 10).

Group	Frequency	Percentage
0	105	14.36%
1	105	14.36%
2	104	14.23%
3	104	14.23%
4	104	14.23%
5	104	14.23%
6	105	14.36%

Table 2: Distribution of `weekday`

Figure 10: Distribution of days of the week



Identify the test to be used (justify your choice)

We have a normally distributed outcome variable `cnt` and a categorical variable of more than 2 independent groups `weekday`. Bartlett's test was conducted to investigate the homogeneity of variance. The null hypothesis is that the variances in groups are equal, so to assume homogeneity we would expect the probability to not be statistically significant. The p-value is 0.378, which is over 0.05, meaning that variances are homogenous in groups. We will use the parametric One-way ANOVA test with `var.equal = TRUE` to determine if the total number of bikes hired per day varies by the day of the week. If a statistically significant difference is found, then we will use Tukey as the post-hoc test.

Conduct the test and report your findings

A one-way between-groups analysis of variance (ANOVA) was conducted to explore the impact of the day of the week on the total number of bikes hired per day. No statistically significant difference was found at the $p < .05$ level ($F(6, 724) = .78$, $p = .58$). We will not continue with a post-hoc test. The means and the standard variations for each group are shown in Table 3 below. Small effect size was also indicated by the Eta squared value (0.01).

Group	Mean	SD
0	4228.83	1872.50
1	4338.12	1793.07
2	4510.66	1826.91
3	4548.54	2038.10
4	4667.26	1939.43
5	4690.29	1874.62
6	4550.54	2196.69

Table 3: Descriptive statistics of `weekday` by group

Based on these findings we will not reject the null hypothesis (H_0) that the total number of bikes hired per day does not vary by the day of the week.

5. Is the weather situation related to the season?

State a hypothesis pair for the question

- Null hypothesis (H_0):
  The weather situation is not related to the season.
- Alternative hypothesis (H_1):
  The weather situation is related to the season.

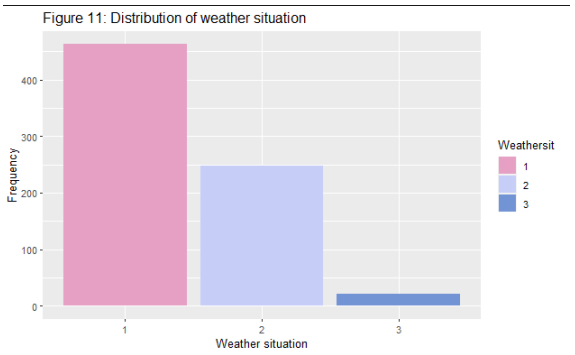
Explore the data to be used to test the hypothesis

Outcome variable: `weathersit`

`weathersit` represents the weather of the day in different types. It is a nominal categorical variable since one type of weather cannot be ranked over or under another. Its groups are independent of each other. The sample size is 731, and there is no missing data. Possible values according to the dataset description are 1, 2, 3, and 4, but the dataset only includes 1, 2, and 3. The most frequently occurring value is 1, making up 63.34% of the observations (see Table 4 and Figure 11).

Group	Frequency	Percentage
1	463	63.34%
2	247	33.79%
3	21	2.87%

Table 4: Distribution of `weathersit`

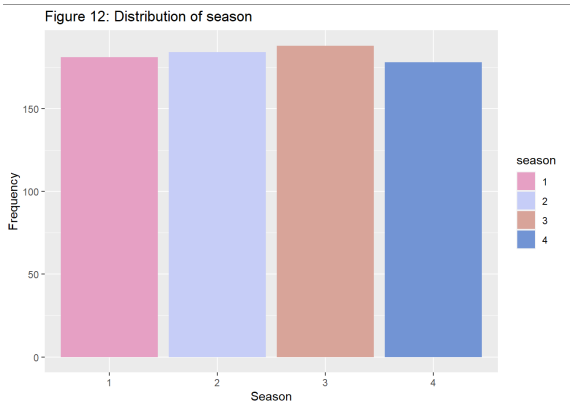


Independent variable: `season`

`season` represents the season when the observation was recorded. It is a nominal categorical variable since one season cannot be ranked over or under another. Its groups are independent of each other. The sample size is 731, and there is no missing data. Possible values are 1, 2, 3, and 4. The values are approximately equally distributed, with the most frequently occurring value 3 making up 25.72% of the observations (see Table 5 and Figure 12).

Group	Frequency	Percentage
1	181	24.76%
2	184	25.17%
3	188	25.72%
4	178	24.35%

Table 5: Distribution of `season`



Identify the test to be used (justify your choice)

We have a categorical outcome variable `weathersit` and a categorical variable of more than 2 independent groups `season`. We will use the Chi-square test to determine if the weather situation is related to the season.

Conduct the test and report your findings

A Chi-square test for independence indicated a statistically significant association between the weather situations and the seasons at $p < 0.05$ level, $\chi^2(6, n = 731) = 14.88, p = 0.021, V = 0.10)$.

Based on this we will reject the null hypothesis (H_0) that the weather situation is not related to the season.

6. Summary

Question 1:
A strong positive relationship was found between the temperature (actual) and the total number of bikes hired per day. While both variables were treated as normal, and the result was found to be statistically significant, it is worth pointing out that the temperature (actual) had an excess kurtosis of -6.17, and the total number of bikes hired per day had an excess kurtosis of -4.48. An investigation of this particular aspect of the data might be warranted.

Question 2:
A weak negative correlation was found between the level of humidity and the total number of bikes hired per day. While both variables were treated as normal, and the result was found to be statistically significant, it is worth pointing out that the total number of bikes hired per day had an excess kurtosis of -4.48. An investigation of this particular aspect of the data might be warranted.

Question 3:
No relationship was found between the total number of bikes hired per day and whether a day is a regular weekday or a weekend. While the total number of bikes hired per day was treated as normal, it is worth pointing out that it had an excess kurtosis of -4.48. An investigation of this particular aspect of the data might be warranted.

Question 4:
No relationship was found between the total number of bikes hired per day and the day of the week. While the total number of bikes hired per day was treated as normal, it is worth pointing out that it had an excess kurtosis of -4.48. An investigation of this particular aspect of the data might be warranted.

Question 5:
A relationship was found between the weather situation and the season.

Works Cited

- Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed., Taylor & Francis, 2013. *University of Toronto Faculty of Arts & Science*, https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf. Accessed 6 November 2022.
- Peck, Roxy, and Jay L. Devore. *Statistics: The Exploration & Analysis of Data*. 7th ed., Cengage Learning, 2012.