Statistics and Data Analysis, Fall 2015
Homework 3: Statistical Estimation and Hypothesis Testing

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

This homework is due ***6:35 pm, November 30, 2015***. Each student should submit her/his own ***hard copy*** to the instructor at the beginning of the class. All the data for this homework are contained in the file "SDA-Fa15_hw03_data.xlsx". Discuss with your classmates but NEVER copy one's work.

Consider the worksheet "Da-An," which contains related information about the YouBike sites in the Da-An District, Taipei, Taiwan. For each site, we collect the number of available bikes (and other information) once per hour for around 31 days. Each site generates 743 rows. As there are 30 distinct sites in the district, there are 22290 rows in total. The columns and their meanings are described below:

- fetch_time: time of fetching the data.
- month: the month part of the fetching time.
- day: the day part of the fetching time.
- weekday: 1 if the day is a weekday and 0 otherwise.
- hour: the hour part of the fetching time.
- station_id: station identification number.
- total_parking_num: number of parking racks of a station.
- available_bike_num: number of available bikes of a station.
- empty_parking_num: number of empty parking of a station.
- latitude: latitude of a station.
- longitude: longitude of a station.
- location: location of a station.
- station_name: name of a station.
- station_address: address of a station.
- in_service: 0 for out of service; 1 for in service.
- weather_type: "Rain" or "No Rain."
- temp: temperature in Kelvin degree.
- pressure: atmospheric pressure in hPa.
- humidity: humidity in percentage.
- wind_speed: wind speed in meter per second.

Just to make grading easier, for all the problems below, please construct your answers with ALL the data given to you. DO NOT try to remove any potential outliers.

1. (10 points) For site i, $i = 1, ..., 30$, find its average number of available bikes between 8 am and 9 am as x_i and that between 9 am and 10 am as y_i. Draw a scatter plot for $\{(x_i, y_i)\}_{i=1,...,30}$ and do some interpretations.

2. (10 points) Given a site, we are interested in factors affecting the number of available bikes. The time in a day is obviously an important factor. To see this, let's take the site "MRT Gongguan Sta.(Exit 2)" as an example. We divide a day into 24 hours, 0 am to 1 am, 1 am to 2 am, ..., and 23 pm to 0 am. For each hour, we calculate the average number of available bikes among all days. For example, for 12 pm to 1 pm, these values are 7 for September 7, 22 for September 8, ..., and 13 for October 7. The average is 12.48. Repeating this process for all the other 23 hours, we may get all the 24 averages. They are presented in a line chart in Figure **??**.

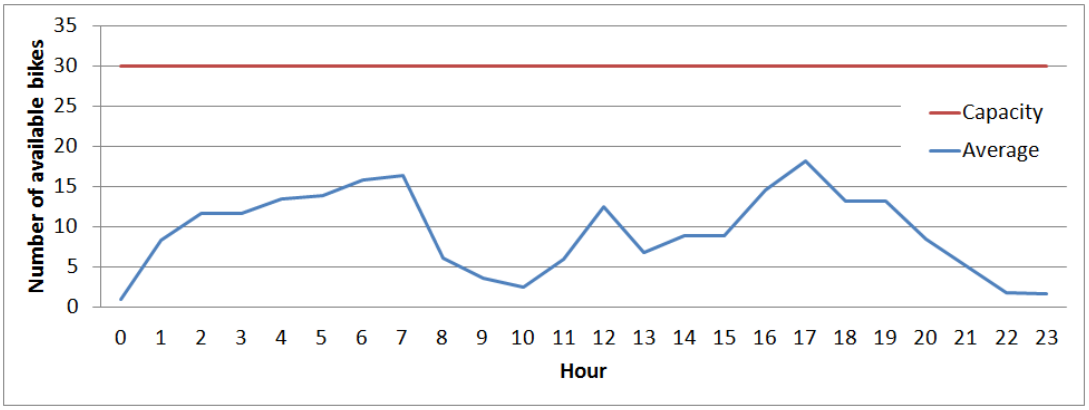

Figure 1: Hourly average number of available bikes in "MRT Gongguan Sta.(Exit 2)"

Please consider the other site "Roosevelt & Xinsheng S. Intersection" and do the same thing to draw a similar line chart.

3. (20 points) It seems to us that the hourly pattern may be different between weekdays and weekends. Therefore, we separate the two types of days and calculate the hourly averages for weekdays and weekends, respectively. For the site "MRT Gongguan Sta.(Exit 2)," the line charts for the two series of average hourly available bikes are depicted in Figure **??**. We can see that the two lines are indeed different. In particular, for weekdays, the availability drops significantly between 8 am to 9 am, which may be due to the fact that many people take MRT to Gongguan and then ride YouBikes to their destinations. This effect is not so strong for weekends.

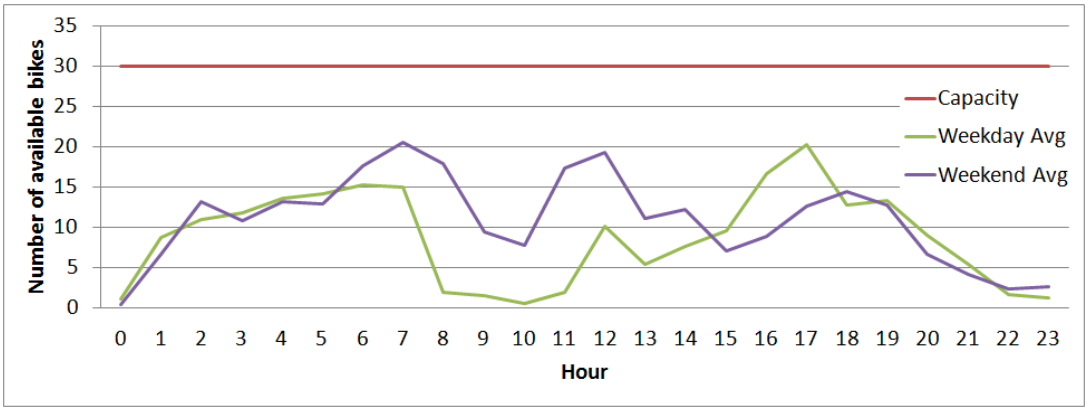

Figure 2: Hourly average number of available bikes during weekdays and weekends in "MRT Gongguan Sta.(Exit 2)"

(a) (10 points) Please consider the other site "N.T.U.S.T" and do the same thing to draw two similar line charts.

(b) (5 points) Identify at least one big drop or jump for the weekday line you depict in Part (a). Try to explain why that drop or jump exists.

Note. The way of providing explanations in this problem has nothing to do with Statistics. This problem is just to help you feel the power of data analysis and visualization. Also, this is an open question and has no standard answer.

(c) (5 points) Identify at least one difference between the weekday and weekend lines you depict in Part (a). Try to explain why that difference exists.

4. (25 points) Let's ignore the difference between weekdays and weekends and consider all days altogether. Given a time and a site, we are always interested in knowing whether the "stocking level" of available bikes is neither too low nor too high. We say that the stocking level is "good" if the number of available bikes is between 20% and 80% of the capacity of a site. For example, the number of available bikes should be within 6 and 24 for the site "MRT Gongguan Sta.(Exit 2)," whose capacity is 30, to be considered to be good.

   For a given site and a given hour, we know the average number of available bikes during the one-month period. However, this is just a sample mean rather than a population mean. We want to estimate the population mean to see if the population mean is within 20% and 80% of the capacity. Therefore, we should construct a confidence interval for the population mean. If the whole confidence interval lie within 20% and 80% of the capacity, we may then be confident (at a given confidence level) that the population mean can be considered good.

   The 95% confidence intervals for all hours have been constructed for the site "MRT Gongguan Sta.(Exit 2)." They are plotted in Figure **??**. We can see that, for example, the stocking level is quite good between 2 am to 8 am. Nevertheless, the lower bound of the confidence interval drops to be lower than 6 for the hour 8 am to 9 am. We thus do not say that the stocking level in the hour 8 am to 9 am is good, even though the sample mean is 6.06, which is above 6.

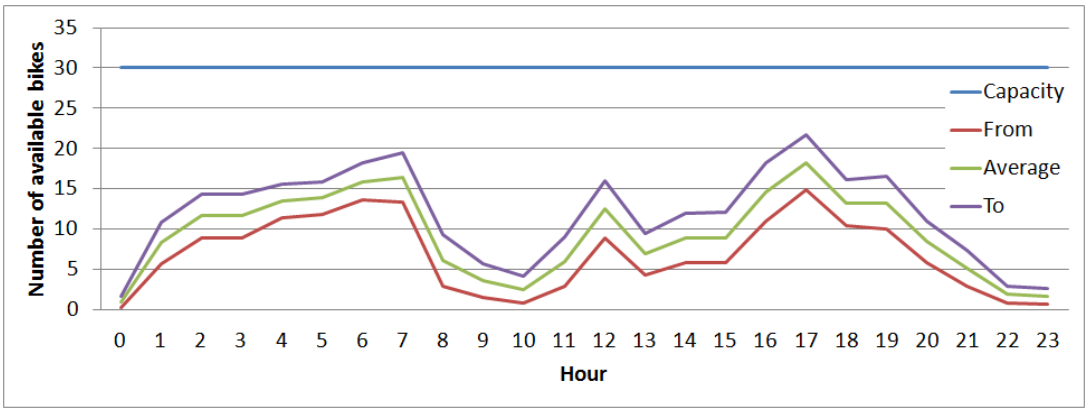

Figure 3: Confidence intervals for hourly average number of available bikes in "MRT Gongguan Sta.(Exit 2)." "From" and "To" are the lower and upper bounds of a confidence interval

Please consider the other site "NTU Information Bldg." to answer the following questions.

(a) (10 points) Find the sample standard deviations for all hours and depict them by a line chart.

(b) (10 points) Construct the 95% confidence intervals for all the hours. Then draw a figure that is similar to Figure **??** to illustrate these confidence intervals. Clearly indicate which distribution is adopted and why that is appropriate.

(c) (5 points) Find those hours whose stocking level is not good, i.e., those hours that we are not 95% confident that the population mean is within 20% and 80% of the capacity.

5. (35 points) Recall that in average we only have 6.06 bikes available at "MRT Gongguan Sta.(Exit 2)" between 8 am and 9 am. As such a low availability can discourage potential customers to include YouBike in their commuting plans, a YouBike manager decides to hire some workers to stay at sites to "control the inventory." For each of these sites, many YouBikes will be locked around the site but not on the parking racks. A worker can unlock these bikes when the availability is low and

"replenish" these bikes to the racks to make them available. The question is to identify several sites to put workers.[1]

(a) (5 points) For all the sites, calculate the average number of available bikes between 8 am and 9 am. Then divide these numbers by site capacities to get the availability rates. Identify those sites whose availability rates are lower than 30%.[2]

(b) (10 points) Knowing that the available numbers of bikes calculated based on the given data are just sample means, the manager wants to test whether the population means of these sites are indeed lower than 30% of the capacities. She first focuses on "MRT Gongguan Sta.(Exit 2)" and ask you to do her a favor. Conduct hypothesis testing to test whether the population mean (i.e., the average number of available bikes between 8 am and 9 am *for all days*) is lower than 9, 30% of its capacity. The manager will put a worker there only if the population mean is that low. Determine whether a worker should be allocated to this site.[3]

Hint. As we do not have the population variance, please use the sample variance instead. As the sample size is above 30, a z test can be used.

(c) (10 points) For each of the 30 sites, repeat the testing in Part (b). Identify those sites that should be allocated workers.

(d) (10 points) Compare the list you compile in Part (a) (with only sample means) and Part (c) (with tests with respect to population means). Which list is longer? Why? Is one list a subset of the other list? Is that always the case or just a coincidence? Please explain.

6. (Bonus: 20 points) Weather naturally affects people's willingness of renting YouBikes. Intuitively, when it is raining, fewer people would be riding bikes. This should leave more YouBikes available on the racks. Figure **??**, which depicts the average number of available bikes when it is rainy or not, confirms the intuition.

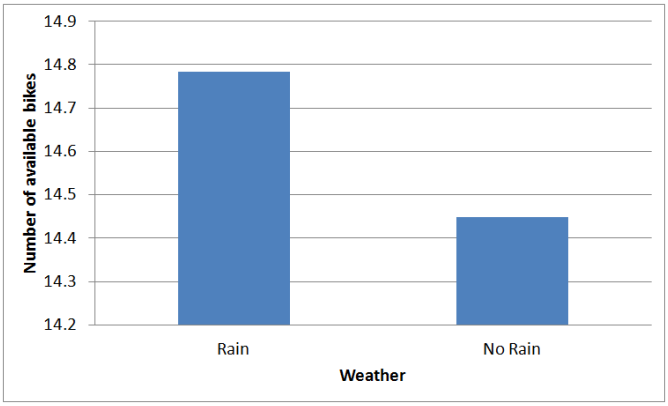

Figure 4: Average number of available bikes under different weather conditions"

Some people believe that the above analysis can be done in a more thorough way. In particular, other factors should be considered for us to fully understand the impact of weather. For example, maybe the situation is opposite at some sites? Please try your best to identify other factors that have significant impacts on how the weather affects the availability. Use summarization, visualization, estimation, and/or testing to support your conclusions.

[1] We all agree that the data provided to you for this homework *has already been* affected in practice by the policy described in this problem. Therefore, the data is somewhat biased. Nevertheless, given that it is too difficult to obtain the unbiased data, please go ahead to do the practices just with the data given to you.

[2] The capacity threshold is intentionally set to be 20% in the previous problem and 30% in this problem.

[3] We all agree that a better indicator for the availability may be the *proportion of days* that the number of available bikes between 8 am and 9 am is above 30% of its capacity. Finding the sample proportion and testing the population proportion can also be done. We choose to work with sample mean and population mean to make the problem easier.