Yilang Karen Kang
ARCH 607
Professor Jae-Pil Choi
December 3, 2017

Architectural Research Methodology

Statistical Analysis Exercise

First of all, in order to test the hypotheses given for this exercise, I followed the suggested steps from our test book 'Exploring Research (Salkind, 2013)' to build testing structures, which is as follows:

1. *State the Null Hypothesis*
2. *Establish the Level of Risk (Level of Significance)*
3. *Select the appropriate Test Statistics*
4. *Determine the Critical Value needed for rejection of the Null Hypothesis*
5. *Compare the obtained values to the Critical Value*
6. *If the obtained value does not exceed the Critical Value then, the Null Hypothesis is the most attractive explanation*

(If $p \leq a$ reject the null hypothesis. If $p > a$ fail to reject the null hypothesis.)

Hypothesis 1: Male drivers are more likely to come to a full stop at the stop line (in other words, observe the stop line) than female drivers.

H0 = There is no relationship between gender and observing the stop line.
H1 = There is a relationship between gender and observing the stop line

In social science and related fields, P-Value at 0.05 is chosen conventionally for the Level of Significance and, hence, .05 will be used throughout this exercise unless modifications are necessary to further the test. Since Chi-Square test is a tool to test the probability of independence and to analyze categorical data that are independent groups, Chi-Square test was

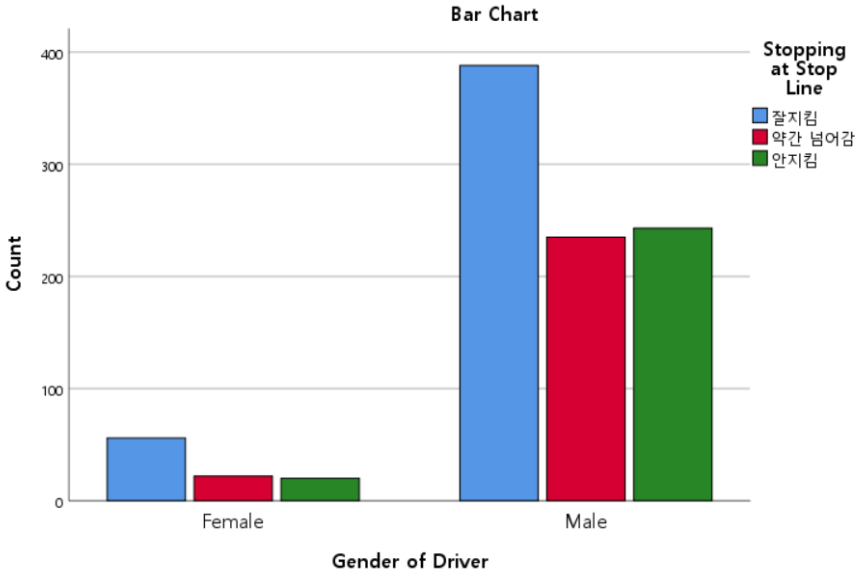
found to be suitable for testing this hypothesis and to find correlation between gender and the driving behaviour. SPSS is used to test the hypotheses and missing values were properly adjusted to provide more accurate statistical analyses.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender of Driver * Stopping at Stop Line	964	96.5%	35	3.5%	999	100.0%

Gender of Driver * Stopping at Stop Line Crosstabulation

			Stopping at Stop Line			
			잘지킴	약간 넘어감	안지킴	Total
Gender of Driver	Female	Count	56	22	20	98
		Expected Count	45.1	26.1	26.7	98.0
		% within Gender of Driver	57.1%	22.4%	20.4%	100.0%
		% of Total	5.8%	2.3%	2.1%	10.2%
	Male	Count	388	235	243	866
		Expected Count	398.9	230.9	236.3	866.0
		% within Gender of Driver	44.8%	27.1%	28.1%	100.0%
		% of Total	40.2%	24.4%	25.2%	89.8%
Total		Count	444	257	263	964
		Expected Count	444.0	257.0	263.0	964.0
		% within Gender of Driver	46.1%	26.7%	27.3%	100.0%
		% of Total	46.1%	26.7%	27.3%	100.0%



Chi-Square Tests

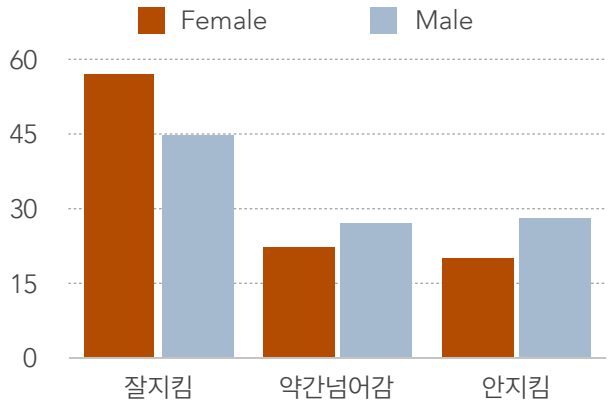
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	5.525[a]	2	.063
Likelihood Ratio	5.539	2	.063
Linear-by-Linear Association	5.034	1	.025
N of Valid Cases	964		

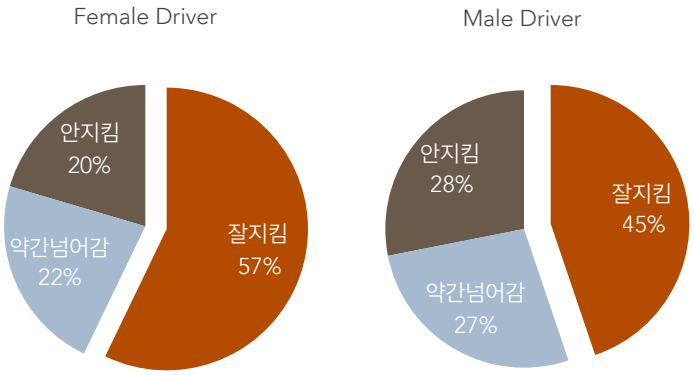
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 26.13.

Tables and Bar Chart Above Show the Results of Chi-Test Between Gender and Observing Stop Line Using SPSS

In the Chi-Square Tests chart above, the highlighted value shows $\chi^2=5.525$, $p=.063$ is greater than Level of Significance=.05. Therefore, the null hypothesis cannot be rejected. In other words, there appears to be no statistically significant relationship between gender and

observing the stop line. Though the bar chart and pie charts below show differences between male and female drivers in observing the stop line (45% and 57%, respectively), the Chi-Square Test results indicate that it is likely due to chance. Therefore, the original hypothesis that male drivers are likely to come to a full stop at the stop line than female drivers cannot be tested with the given samples.





Bar chart and Pie Charts Showing the Percentage of 'Observing Stop Line' Counts (잘지킴, 약간 넘어감, 안지킴) Within Each Gender Group

Hypothesis 2: The older drivers are more likely to observe the stop line than the younger drivers

H0 = There is no relationship between age and observing the stop line.

H1 = There is a relationship between age and observing the stop line.

Similar to the first hypothesis, I started off by verifying if there is a statistically significant relationship between age and driving behaviour (observing the stop line). Chi-square test using SPSS is employed and if

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	23.392[a]	6	.001
Likelihood Ratio	26.384	6	.000
Linear-by-Linear Association	13.270	1	.000
N of Valid Cases	968		

a. 3 cells (25.0%) have expected count less than 5. The minimum expected count is 1.34.

the result rejects H0, then necessary tests will follow to measure how likely the older drivers are to observe the stop line.

However, as shown in the Chi-Square Results tables, the number counts for '40s' category is 5 and the caption below Chi-Square Tests table indicates that 3 cells have expected count less than 5, which makes up for 25%. Basic requirements for Chi-Square test is that each expected counts must be greater than 5 and those with cell count smaller than 5 must be less than 25% of the total count. In order to be more conservative with the test, Fisher's Exact Test was chosen for this test.

Age of Driver * Stopping at Stop Line Crosstabulation

			Stopping at Stop Line			
			잘지킴	약간 넘어감	안지킴	Total
Age of Driver	20대	Count	213	98	111	422
		Expected Count	192.7	112.9	116.4	422.0
		% within Age of Driver	50.5%	23.2%	26.3%	100.0%
		% of Total	22.0%	10.1%	11.5%	43.6%
	30대	Count	220	146	135	501
		Expected Count	228.8	134.0	138.2	501.0
		% within Age of Driver	43.9%	29.1%	26.9%	100.0%
		% of Total	22.7%	15.1%	13.9%	51.8%
	40대	Count	3	0	2	5
		Expected Count	2.3	1.3	1.4	5.0
		% within Age of Driver	60.0%	0.0%	40.0%	100.0%
		% of Total	0.3%	0.0%	0.2%	0.5%
	50대	Count	6	15	19	40
		Expected Count	18.3	10.7	11.0	40.0
		% within Age of Driver	15.0%	37.5%	47.5%	100.0%
		% of Total	0.6%	1.5%	2.0%	4.1%
Total		Count	442	259	267	968
		Expected Count	442.0	259.0	267.0	968.0
		% within Age of Driver	45.7%	26.8%	27.6%	100.0%
		% of Total	45.7%	26.8%	27.6%	100.0%

Chi-Square Result Tables for Age and Observing Stop Line

Chi-Square Tests

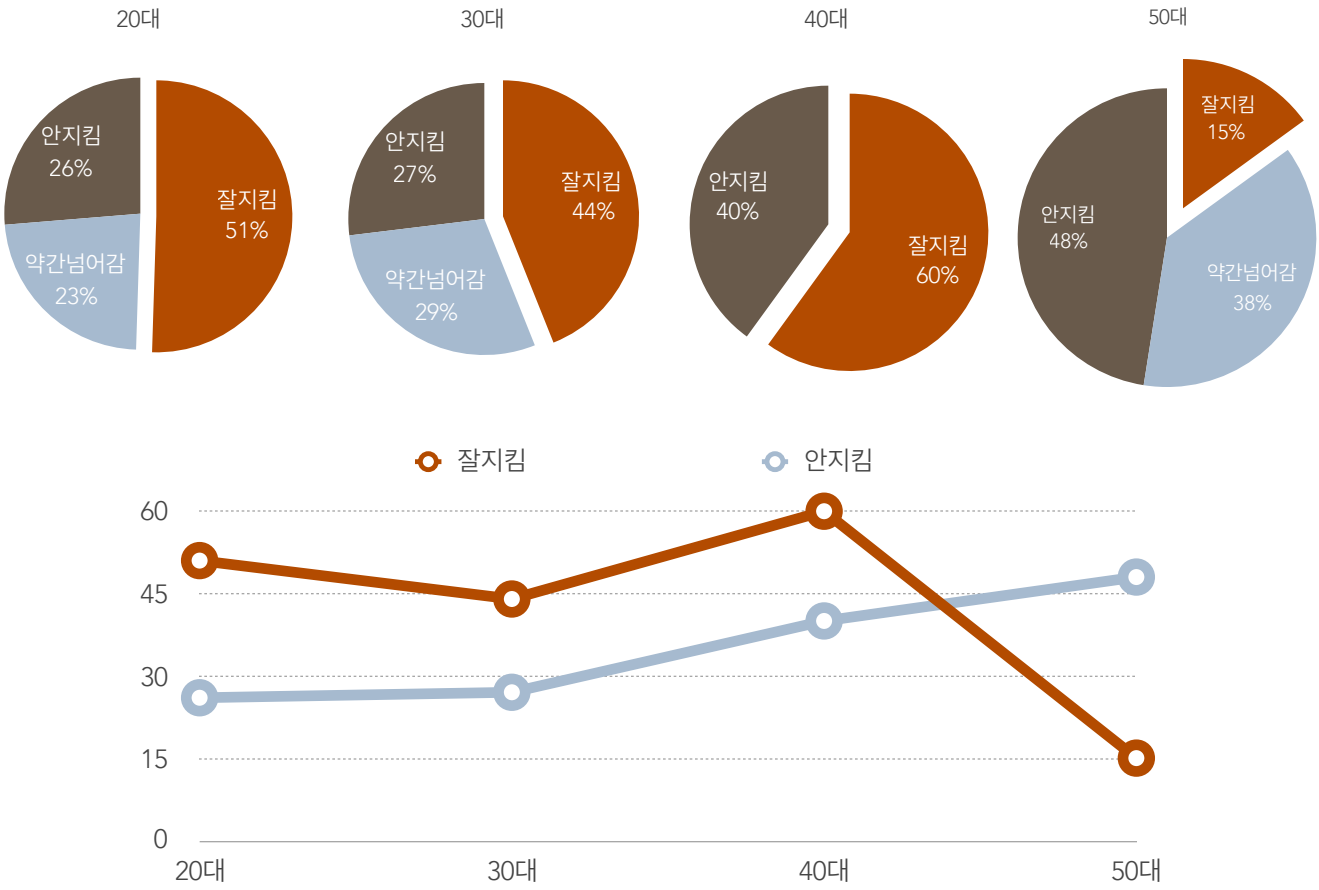
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	23.392[a]	6	.001	.[b]		
Likelihood Ratio	26.384	6	.000	.000		
Fisher's Exact Test	24.755			.000		
Linear-by-Linear Association	13.270[c]	1	.000	.000	.000	.000
N of Valid Cases	968					

a. 3 cells (25.0%) have expected count less than 5. The minimum expected count is 1.34.

b. Cannot be computed because there is insufficient memory.

c. The standardized statistic is 3.643.

The Fisher's Exact Test resulted in P value far less than 0.000..(χ^2=24.755), which is far smaller than the Level of Significance =0.05, hence, the test result rejects H0. In other words, there is a statistically significant relationship between age and observing the stop line. As for the original hypothesis that older drivers are likely to come to a full stop than younger drivers, a new set of graphs were created to compare the percentage of 'Full Stop (잘지킴)' within each group.



Even though the '40s' sample cannot be trusted entirely due to its size, discernible patterns are observed in the pie charts and line chart above. As the age group gets older, the survey rate for 'full stop' declines, though subtle. I also looked at 'not stop' rate differences to support the findings and it is rather clearer that as the age group gets older, the survey rate for 'not stop (안지킴)' increases, suggesting that the original hypothesis that older drivers are likely to observe the stop line is false.

Hypothesis 3: Observing the stop line depends on the color of vehicles.

H0 = There is no relationship between the colour of vehicles and observing the stop line
H1 = There is a relationship between the colour of vehicles and observing the stop line

This test is similar to the first one and first the relationship between the colour of vehicles and observing the stop line needed to be tested using Chi-Square Test.

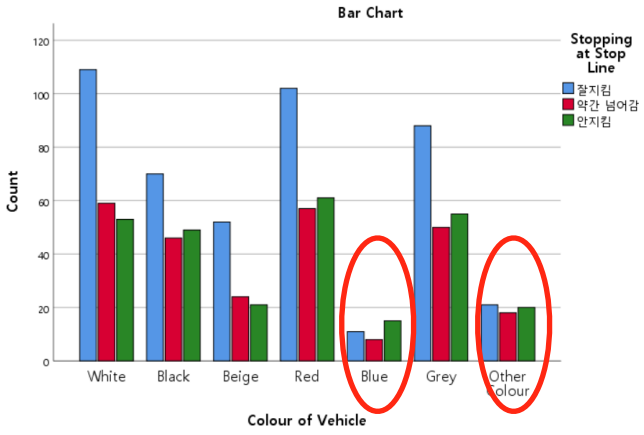
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	12.554[a]	12	.402
Likelihood Ratio	12.327	12	.420
Linear-by-Linear Association	2.791	1	.095
N of Valid Cases	989		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.01.

The result shows χ^2=12.554, p=.402, which is greater than .05 (LOS). Hence, H0 cannot be rejected, meaning that 'the colour of vehicles' is not related to 'observing the stop line' sample. Therefore, the original hypothesis whether or not observing the stop line depends on the colour of vehicle cannot be tested.

However, there is one thing that needs to be taken into consideration is the sample size differences. In more profound studies, it is suggested that sample sizes should be adjusted with probability proportional to the size of each group in order to minimize any potential statistical errors. In this hypotheses testing, it could have been useful and more appropriate to have done so since the sample groups have size differences to a notable extent that may have affected Chi-Square test results.



Hypothesis 4: There are differences in observing stop line among these four groups: older males, younger males, older females, and younger females.

This hypothesis attempts to look at multiple independent variables and their influences on observing the stop line. In order to test the hypothesis, a new table was created to combine age

group and gender. The age group was restructured to have two categories (old and young), instead of 4 from the original data set. The 'old' is a combination of 40s and 50s and 'young' a combination of 20s and 30s. Then, the gender group was combined with the binary age group data.

Table 1

		Gender of Driver					
		Female			Male		
		Age by Young and Old			Age by Young and Old		
		Young Count	Old Count	Total Count	Young Count	Old Count	Total Count
Stopping at Stop Line	잘지킴	56	0	56	375	7	382
	약간 넘어감	21	1	22	222	10	232
	안지킴	20	0	20	224	11	235
	Total	97	1	98	821	28	849

Pearson Chi-Square Tests

		Gender of Driver	
		Female	Male
		Age by Young and Old	Age by Young and Old
Stopping at Stop Line	Chi-square	3.490	4.727
	df	2	2
	Sig.	.175[a,b]	.094

Results are based on nonempty rows and columns in each innermost subtable.

a. More than 20% of cells in this subtable have expected cell counts less than 5. Chi-square results may be invalid.

b. The minimum expected cell count in this subtable is less than one. Chi-square results may be invalid.

The H0 here is 'There is no relationship between Gender_Age and Observing the Stop Line'. According to the Chi-Square Test, both P-values are greater than .05. Hence, there are no statistically significant relationships between the Gender_Age group and driving behaviour. The table below was created to see the percentage differences of each variable, in which the sample sizes of old and young in each gender group are notably different. This test again will be improved with more data collected for 'Female_Old' and 'Male_Old' for more accurate testing of the original hypothesis.

	Female_Young	Female_Old	Female Total	Male_Young	Male_Old	Male Total
잘지킴	57%	0%	57%	44%	1%	45%
약간넘어감	21%	1%	22%	26%	1%	27%
안지킴	21%	0%	21%	26%	2%	28%
Total			100%			100%

Hypothesis 5: The vehicles from Seoul are more likely to observe the stop line than those of Kyonggi Province.

H0 = There is no relationship between the origin of car registry and observing the stop line

H1 = There is a relationship between the origin of car registry and observing the stop line

The same method of test applies here. Chi-Square Test was applied to test the relationship. The result shows χ^2=6.185, p=.186, which is greater than .05 (LOS). Therefore, the null hypothesis cannot be rejected and there is no statistically significant relationship between observing the stop line and the origin of vehicle registry.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	6.185[a]	4	.186
Likelihood Ratio	6.017	4	.198
Linear-by-Linear Association	3.225	1	.073
N of Valid Cases	967		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.20.

Chi-Square Tests

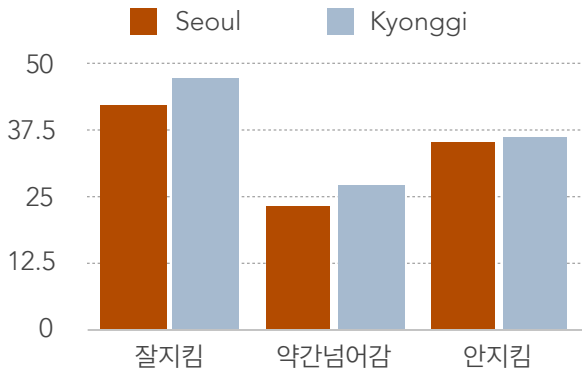
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	6.179[a]	2	.046
Likelihood Ratio	6.011	2	.050
Linear-by-Linear Association	4.261	1	.039
N of Valid Cases	928		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 57.35.

Chi-Square Tests (Left) - for Seoul, Kyonggi and Other
Chi-Square Tests (Right) - for Seoul and Kyonggi

However, when the test only between Seoul and Kyonggi samples were compared, it resulted in different statistical significance as shown in the chart above (on the right). The P-value was lower than .05 and it rejects H0. There is a statistical significance between region and observing the stop line.

A bar graph was created based on the Chi-Square results and Kyonggi vehicles are slightly better at observing the stop line than Seoul vehicles, rejecting the original hypothesis.



Hypothesis 6: Domestic car drivers are more likely to observe the stop line than foreign car drivers.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	.414[a]	2	.813
Likelihood Ratio	.415	2	.813
Linear-by-Linear Association	.084	1	.771
N of Valid Cases	845		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.40.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Car Makes * Stopping at Stop Line	845	84.6%	154	15.4%	999	100.0%

Car Makes * Stopping at Stop Line Crosstabulation

			Stopping at Stop Line			
			갖지킴	약간 넘어감	안지킴	Total
Car Makes	Domestic	Count	373	207	214	794
		Expected Count	373.0	208.6	212.4	794.0
		% within Car Makes	47.0%	26.1%	27.0%	100.0%
		% of Total	44.1%	24.5%	25.3%	94.0%
	Foreign	Count	24	15	12	51
		Expected Count	24.0	13.4	13.6	51.0
		% within Car Makes	47.1%	29.4%	23.5%	100.0%
		% of Total	2.8%	1.8%	1.4%	6.0%
Total		Count	397	222	226	845
		Expected Count	397.0	222.0	226.0	845.0
		% within Car Makes	47.0%	26.3%	26.7%	100.0%
		% of Total	47.0%	26.3%	26.7%	100.0%

H0 = There is no relationship between car makes (domestic and foreign) and observing the

stop line

H1 = There is a relationship between car makes and observing the stop line

In terms of statistics, it requires Chi-Square test to test the relationship between the two variables. However, the original data set divides car makes into 5 (현대, 대우, 기아, 쌍용, 외제), consisting of 4 domestic brands and 1 category for that all foreign makes. Since the original hypothesis tries to specifically test the influence of domestic car and foreign car makes on observing the stop line, the date set has been adjusted to divide the sample group into 2 categories--domestic and foreign prior to running Chi-Square Test.

The Chi-Square Tests shows χ^2=.414, p=.813, which is greater than .05 (LOS). Hence, the test fails to reject the null hypothesis. Hence, there is no statistically significant relationship

between car makes and observing the stop line and the original hypothesis is rejected. Note that for the similar test results above, when the Chi-Square tests fail to reject the null hypothesis, it does not mean that these two categorical variables are independent. Rather, it only attests to the fact that there is not enough evidence that they are dependent.

Hypothesis 7: Drivers are less likely to observe the stop line on Friday

To test the hypothesis, first off, the relationship between Day of Survey Taken and Observing the Stop Line.

H0 = There is no relationship between Day of Survey Taken and Observing the Stop Line

H1 = There is a relationship between Day of Survey Taken and Observing the Stop Line

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	30.039[a]	6	.000
Likelihood Ratio	31.204	6	.000
N of Valid Cases	992		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.39.

As shown in the Chi-Square Tests, P-value is less than 0.05, which rejects the null hypothesis. In other words, there is sufficient evidence that there is statistically significant relationship between survey day and observing the stop line.

Although there appears to be a significant relationship between the two variables, the survey day data were not entered with numeric codes. In order to find out if the stopping rate increases or decreases as days pass by during the week, it will require further tests. With the given data set, pie charts below do show Fridays and Wednesdays show lower rates for 'not stopping' compared to Tuesday and Thursday. Provided that the survey day and observing the stop line are dependent, the hypothesis can be partially accepted that drivers indeed are less likely to observe the stop line on Friday along with Wednesday.