เรื่อง การวิเคราะห์ข้อมูลรายรับรายจ่ายและข้อมูลการเดินทาง

จัดทำโดย

นายคเชนทร์ ธรรมมาสถิตย์กุล รหัสประจำตัว 6010502497

เสนอ

ผศ.ดร.สุภาพร เอื้อจงมานี

รายงานนี้เป็นส่วนหนึ่งของการเรียนวิชา

01204314 Statistics for Computer Engineering Applications

มหาวิทยาลัยเกษตรศาสตร

คำนำ

รายงานเล่มนี้เป็นส่วนหนึ่งของวิชา 01204314 Statistics for Computer Engineering Applications โดยมีจุดประสงค์เพื่อคำนวณ

ทั้งนี้ทางผู้จัดทำหวังเป็นอย่างยิ่งว่ารายงานเล่มนี้สามารถเป็นประโยชน์ต่อผู้ที่เข้ามาศึกษาไม่มากก็น้อย และหากมีข้อผิดผลาดประการใด ทางผู้จัดทำต้องขออภัยมา ณ ที่นี้ด้วย

นาย คเชนทร์ ธรรมมาสถิตย์กุล

ผู้จัดทำ

สารบัญ

หัวข้อ หน้า

การเก็บรวบรวมข้อมูล

ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลรายรับรายจ่ายของ นาย คเชนทร์ ธรรมมาสถิตกุล โดยเก็บในช่วงวันที่ 9 มกราคม 2563 ถึง 22 มกราคม 2563 เป็นเวลา 2 อาทิตย์ และ จะเก็บข้อมูลได้แก่

- วันที่
- ทำอะไร
- รายรับ
- รายจ่าย
- คงเหลือ
- เดินทางด้วยอะไร
- จากไหน
- ถึงไหน
- ระยะทาง(โดยประมาณ)

วันที่	ทำอะไร	รายรับ	รายจ่าย	คงเหลือ	เดินทางด้วยอะไร	จากไหน	ถึงไหน	ระยะทาง(โดยประมาณ)
9/1/2563	คงเหลือ	6728	0	6728				
9/1/2563	ข้าวเช้า + น้ำ 1 ขวด	0	47	6681				
9/1/2563	ข้าวเที่ยง	0	25	6656				
9/1/2563	ข้าวเย็น	0	280	6376				
10/1/2563	เต้าหู้ทอด	0	20	6356				
10/1/2563	ข้าวเที่ยง	0	40	6316				
11/1/2563	น้ำเปล่า + ไอติม	0	32	6284				
12/1/2563	ลับ1	0	1230	5054				

และ ข้อมูลค่าใช้จ่ายในการเดินทางของส่วนกลาง จะเก็บข้อมูลอันได้แก่

- Distance
- Expense
- From
- To

Distance	Expense	From	To	ID
1 kilometer	15 Baht	BTS ม.เกษตร	ประตูงาม 3	11
0.55 killometer	10 Bath	งามวงศ์วาน1	ตึกศูนย์เรียนรวม1	13
0.6 killometer	10 Bath	งามวงศ์วาน1	BTS ม.เกษตร	13
0.6 killometer	10 Bath	BTS ม.เกษตร	งามวงศ์วาน1	13

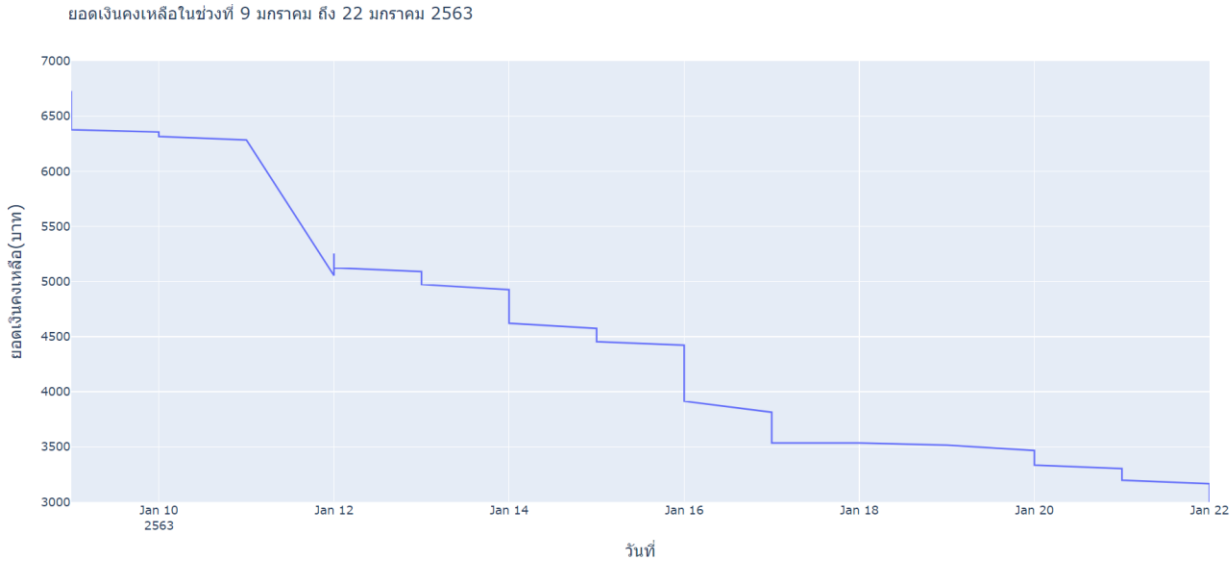
โดยในค่าใช้จ่ายในการเดินทางของส่วนกลางจะมีข้อมูลของมอเตอร์ไซค์รับจ้าง, แท็กซี่, รถประจำทาง, BTS, MRT, รถไฟ

* ข้อมูลรายรับรายจ่ายจะอยู่ในไฟล์ **“รายรับ-จ่าย Khachen Thammasathidkul.xlsx”**

** ข้อมูลค่าใช้จ่ายในการเดินทางของส่วนกลางจะอยู่ในไฟล์ **“Transportation Data.xlsx”**

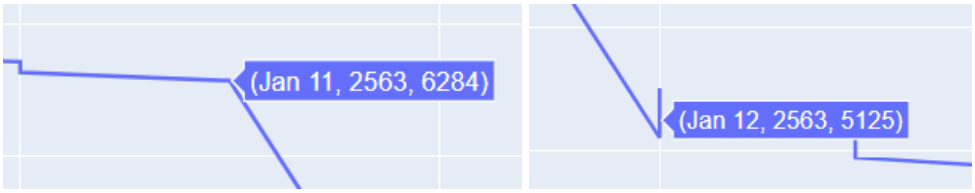
การแสดงผลข้อมูล

1. Time-series

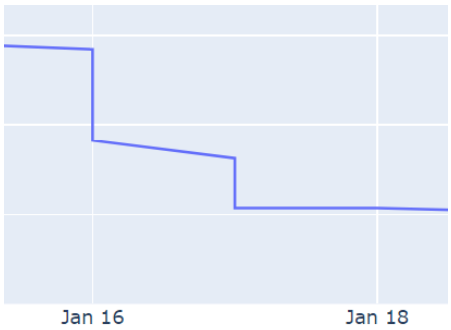


ที่เลือกข้อมูลจำนวนเงินคงเหลือมาทำกราฟ time-series เพราะว่าอยากจะเห็นว่าวันไหนบ้างที่จำนวนเงินของผมลดลงมากเป็นพิเศษ เพราะว่าเดือนที่ผ่านมาเงินในบัญชีของผมลดลงเป็นอย่างมาก จึงอยากจะรู้ว่าวันไหนบ้างที่ทำใช้จ่ายเยอะเป็นพิเศษ

เท่าที่สังเกตกราฟยอดเงินคงเหลือจะสังเกตว่าวันที่ 11 และ วันที่ 12 มกราคม มีจำนวนเงินลดลงอย่างมาก ที่บันทึกจะเป็นเป็นลับ1

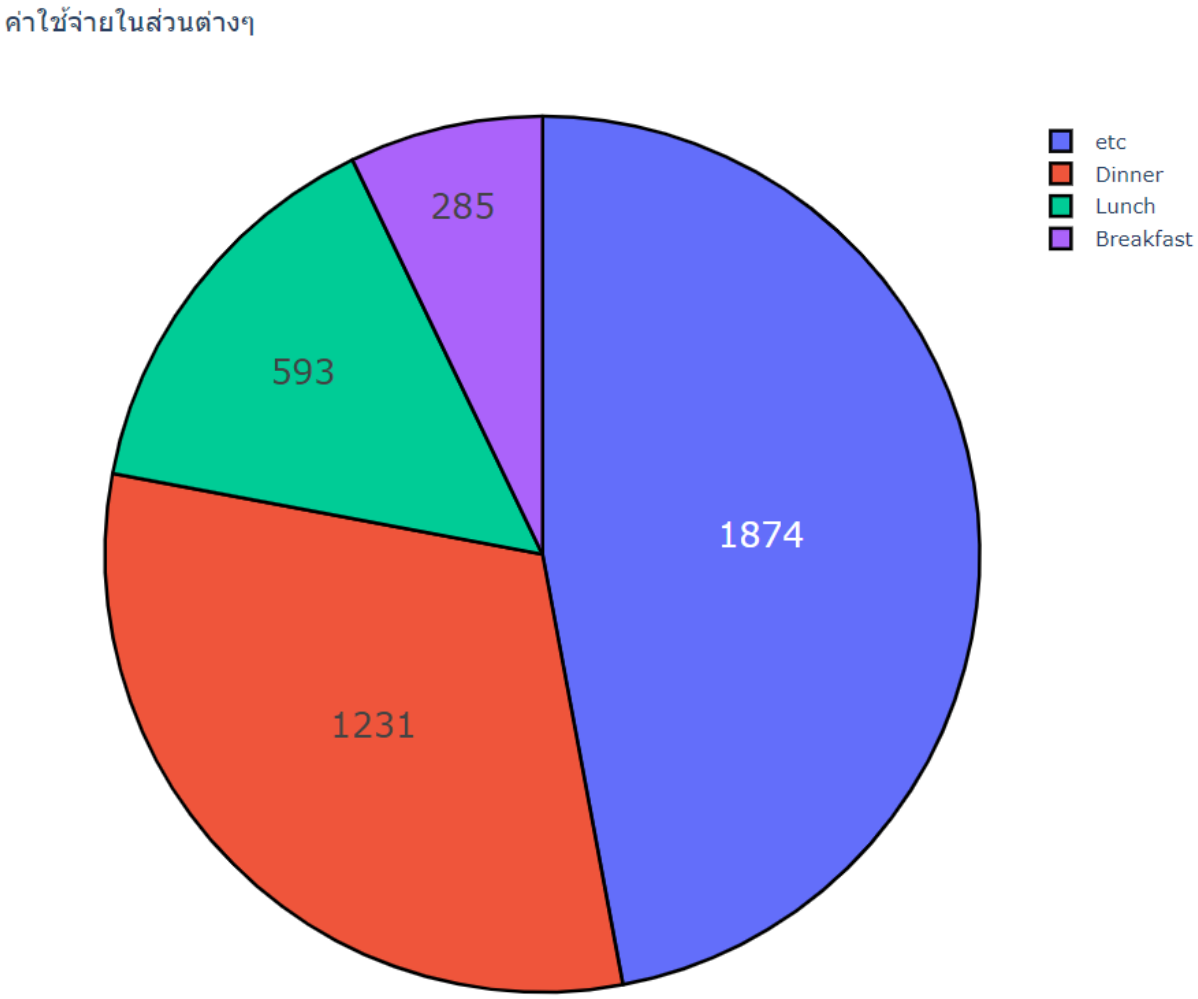


และในวันที่ 16 มกราคม มีจำนวนเงินลดลงเยอะอีกรอบหนึ่ง ซึ่งเป็นลับ2



ทั้งลับ1 และ ลับ2 จะเป็นการเติมเงินเข้าเกม จึงทำให้เงินในบัญชีลดลงอย่างมาก หลังจากที่ผมได้เห็นการลดลงของยอดเงินคงเหลือลดลงอย่างมากขนาดนี้ ในเดือนต่อมาผมจึงลดค่าใช้จ่ายในส่วนนี้จึงทำให้เงินในบัญชีเหลือเยอะขึ้นอย่างมาก

2. Part-to-whole

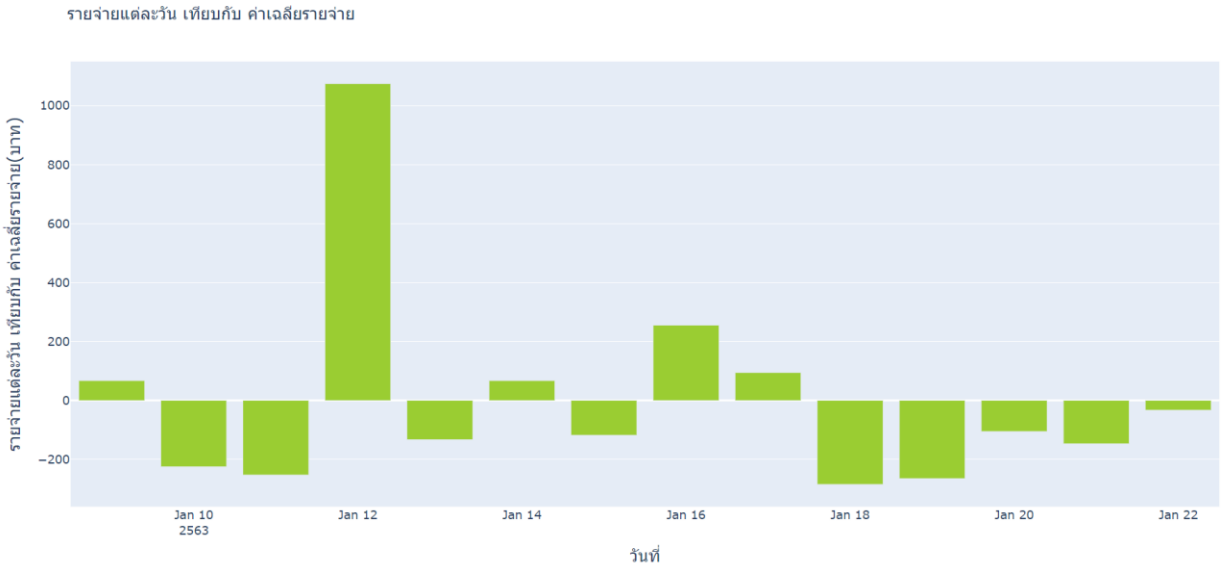


ที่ทำกราฟนี้เป็นเพราะว่าอยากจะรู้ว่าตัวผมเองใช้จ่ายในมื้ออาหารต่างๆ มีสัดส่วนเป็นอย่างไรเพื่อที่จะลดค่าใช้จ่ายได้

จากที่สังเกตจากกราฟ จะเห็นได้ว่าอาหารเย็น มีอัตราส่วนที่มากการมื้ออื่นมากๆ เป็นเพราะว่าเพื่อนชวนไปกินไหนผมไปหมด จึงทำให้เดือนถัดมาลองปฏิเสธในมื้อเย็นที่จะไปกับเพื่อนดูบ้างจะเป็นอย่างไร

และในส่วน etc จะรวมถึง ลับ1 และ ลับ2 ซึ่งเป็นค่าใช้จ่ายที่มากจึงทำให้ดูเป็นชิ้นกราฟที่ใหญ่เป็นพิเศษ การทำกราฟนี้จึงทำให้ควบคุมค่าใช้จ่ายต่างๆได้ดีขึ้น

3. Deviation



กราฟนี้เป็นรายจ่ายแต่ละวันเมื่อเทียบกับค่าเฉลี่ยของรายจ่ายในแต่ละวัน เพื่อที่จะดูว่าวันไหนบ้างมีการใช้จ่ายที่เกินค่าเฉลี่ยมาบ้าง

ก็จะมีวันที่ 9, 12, 14, 16 และ 17 มกราคมที่เกินค่าเฉลี่ยมา โดยเฉพาะอย่างยิ่ง วันที่ 12 และ 16 มกราคมตามที่วิเคราะห์ใน time-series และวันอื่นๆที่เกินค่าเฉลี่ย จะเป็นวันที่เพื่อนชวนไปกินข้าวเย็นซึ่งจะเห็นได้ดังกราฟ

* กราฟอยู่ในไฟล์ **“assignment1.ipynb”**

Hypothesis Test on Two Data Sets

จะทำการหาว่าค่าเฉลี่ยของค่าใช้จ่ายใน 2 อาทิตย์มีค่าเท่ากันหรือไม่ เมื่อแบ่งข้อมูลเป็นขนาด 1 อาทิตย์ จำนวน 2 เซ็ต ดังรูป

	day	spend
0	2563-01-09 00:00:00	352
1	2563-01-10 00:00:00	60
2	2563-01-11 00:00:00	32
3	2563-01-12 00:00:00	1359
4	2563-01-13 00:00:00	152
5	2563-01-14 00:00:00	352
6	2563-01-15 00:00:00	167

	day	spend
7	2563-01-16 00:00:00	540
8	2563-01-17 00:00:00	379
9	2563-01-18 00:00:00	0
10	2563-01-19 00:00:00	20
11	2563-01-20 00:00:00	180
12	2563-01-21 00:00:00	138
13	2563-01-22 00:00:00	252

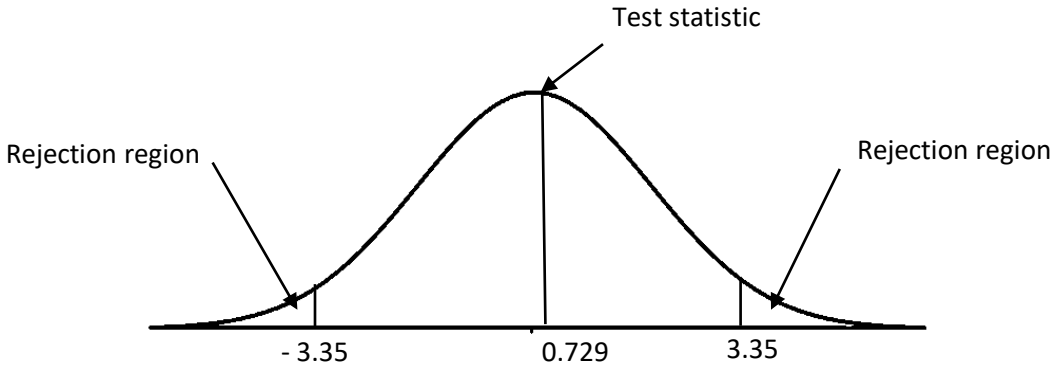
จากการดูข้อมูลจึงเลือก population mean test แบบ normal and small sample size เพราะมีจำนวนข้อมูลน้อย และ เราจะทำการหาค่าเฉลี่ยว่าข้อมูลทั้ง 2 เซ็ต เท่ากันหรือไม่จึงตั้ง null hypothesis และ alternative hypothesis ดังนี้

H_0 : ค่าเฉลี่ยของทั้ง 2 สัปดาห์มีค่าเท่ากัน ($\mu_1 - \mu_2 = 0$)

H_a : ค่าเฉลี่ยของทั้ง 2 สัปดาห์มีค่าไม่เท่ากัน ($\mu_1 - \mu_2 \neq 0$)

จะใช้ significance level อยู่ที่ 0.01

และเมื่อคำนวณค่า test statistic จะได้ค่าเท่ากับ 0.729 และค่า degree of freedom เท่ากับ 8 และเมื่อหาค่า $t_{\alpha/2\,,\,8} = t_{0.005,\,8} = 3.35$



https://www.adelaide.edu.au/mathslearning/resources/statprac1/normal-dist-word.html

จากรูปทางด้านบนจะเห็นว่าค่า Test statistic ไม่ตกอยู่ใน rejection region จึงไม่ปฏิเสธ null hypothesis จึงสรุปได้ว่าค่า เฉลี่ยของทั้ง 2 สัปดาห์ มีค่าเท่ากัน

* ค่าคำนวณต่างๆอยู่ในไฟล์ **“assignment1.ipynb”**

Anova on single factor

การทำข้อนี้ต้องการที่จะรู้ว่าค่าเฉลี่ยของค่าใช้จ่ายในอาหารแต่ละมื้อมีค่าเท่ากันหรือไม่โดยจะทำการปรับตารางดังรูปด้านล่างนี้

	Price(baht)							
Breakfast	47	47	47	32	48	32	32	
Lunch	25	40	35	255	40	20	100	38
dinner	280	105	50	55	139	254	125	48

Null hypothesis และ alternative hypothesis จะได้ดังนี้

H_0 : ค่าเฉลี่ยของทั้ง 3 มื้อมีค่าเท่ากัน ($\mu_1 = \mu_2 = \mu_3$)

H_a : มีค่าเฉลี่ยอย่างน้อยหนึ่งค่าที่มีค่าไม่เท่ากัน

จะใช้ significance level อยู่ที่ 0.01

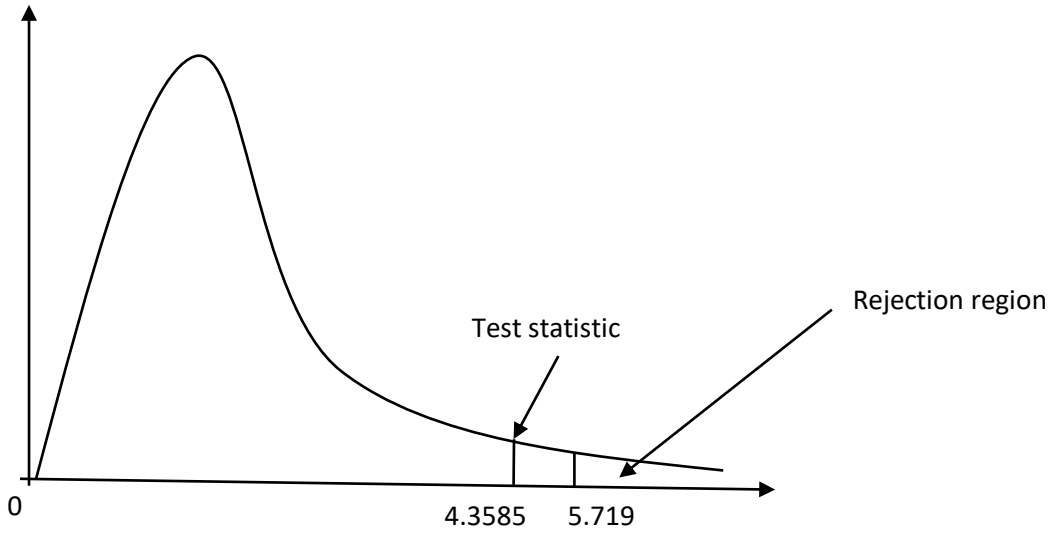
และเมื่อคำนวณค่า

- Degree of freedom : df
- Sum of Square : SS
- Mean Square : MS
- Test statistic : f

จะได้ค่าดังตารางด้านล่างนี้

	df	SS	MS	f
Treatment	2	41133.89	20566.94	4.3585
Error	22	103813.9	4718.812	
Total	24	144947.8		

และจะได้ $F_{0.01,\,2,\,22} = 5.719$



จากรูปทางด้านบนจะเห็นว่าค่า Test statistic ไม่ตกอยู่ใน rejection region จึงไม่ปฏิเสธ null hypothesis จึงสรุปได้ว่าค่า เฉลี่ยของมื้ออาหารทั้ง 3 มื้อมีค่าเท่ากัน

* ค่าคำนวณต่างๆอยู่ในไฟล์ **“food_price_per_meal.xlsx”**

Anova on two factors (additive)

ในข้อนี้ต้องการที่จะหาว่าระหว่างมื้ออาหารและวันส่งผลต่อค่าเฉลี่ยที่ใช้จ่ายในมื้ออาหารหรือไม่ และได้ทำการจัดตารางดังนี้

- มีมื้ออาหารเป็นปัจจัยแรก (ปัจจัยA)
- มีวันในสัปดาห์เป็นปัจจัยที่สอง (ปัจจัยB)
- ค่าในตารางเป็นค่าใช้จ่ายรวมของวันและมื้อนั้นๆ

		Day				
		Moday	Tuesday	Wednesday	Thursday	Friday
Meal	Breakfast	48	79	79	79	0
	Lunch	35	293	80	45	140
	dinner	230	98	230	419	254

Hypothesis บนปัจจัยแรก :

H_{0A} : $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (ปัจจัยA ไม่ส่งผลต่อค่าเฉลี่ย)

H_{aA} : มี α อย่างน้อย 1 ตัวที่มีค่าไม่เท่ากัน (ปัจจัยA ส่งผล)

Hypothesis บนปัจจัยแรก :

H_{0B} : $\beta_1 = \beta_2 = \beta_3 = 0$ (ปัจจัยB ไม่ส่งผลต่อค่าเฉลี่ย)

H_{aB} : มี β อย่างน้อย 1 ตัวที่มีค่าไม่เท่ากัน (ปัจจัยB ส่งผล)

จะใช้ significance level อยู่ที่ 0.05

และเมื่อคำนวณค่า

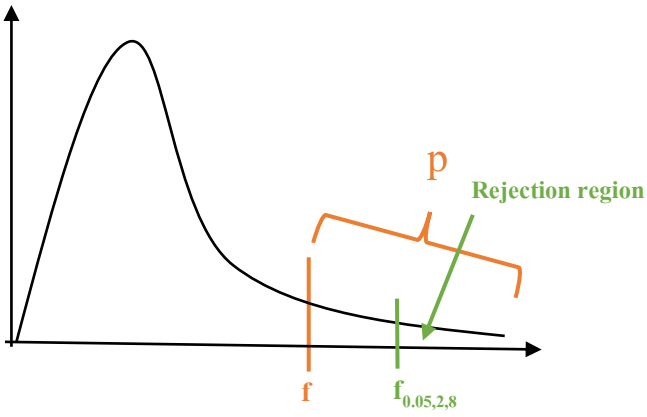
- Degree of freedom : df
- Sum of Square : SS
- Mean Square : MS
- Test statistic : f

จะได้ค่าดังตารางด้านล่างนี้

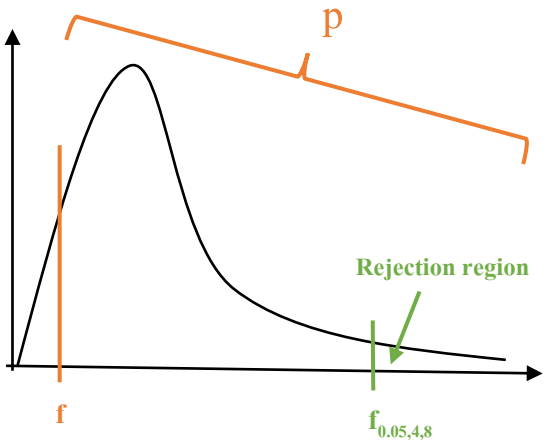
	df	SS	MS	f
A	2	93121.6	46560.8	4.060812
B	4	10232.93	2558.233	0.223117
Error	8	91727.07	11465.88	
Total	14	195081.6		

และเมื่อหาค่า p-value ของ ปัจจัยA และ ปัจจัยB จะได้

p-value ของ A เท่ากับ 0.061



p-value ของ B เท่ากับ 0.918



ไม่ปฏิเสธทั้ง H_{0A} และ H_{0B} เพราะค่า p-value ของ A มีค่าน้อยกว่าค่า significance level และค่า p-value ของ B มีค่าน้อยกว่าค่า significance levelแสดงว่าทั้งมื้ออาหารและวันไม่ส่งผลต่อค่าเฉลี่ยที่ใช้จ่ายในมื้ออาหาร

* ค่าคำนวณต่างๆอยู่ในไฟล์ **“Anova2.xlsx”** และ **“assignment1.ipynb”**

Categorical data analysis (homogeneity)

ในข้อนี้ต้องการที่จะทราบว่าในการเดินทางระหว่าง มอเตอร์ไซค์รับจ้าง และ รถประจำทาง มีอัตราส่วนในการโดยสารในแต่ละช่วงราคาเท่ากันหรือไม่โดยค่าในตารางคือจำนวนครั้งที่โดยสารในแต่ละช่วงราคาดังรูป

		Expense				
		0 <= x <= 10 baht	10 < x <= 20 baht	20 < x <= 30 baht	x > 30 baht	sample size
Times	MotorCycle	10	35	2	4	51
	Bus	10	42	2	3	57
	Total	20	77	4	7	108

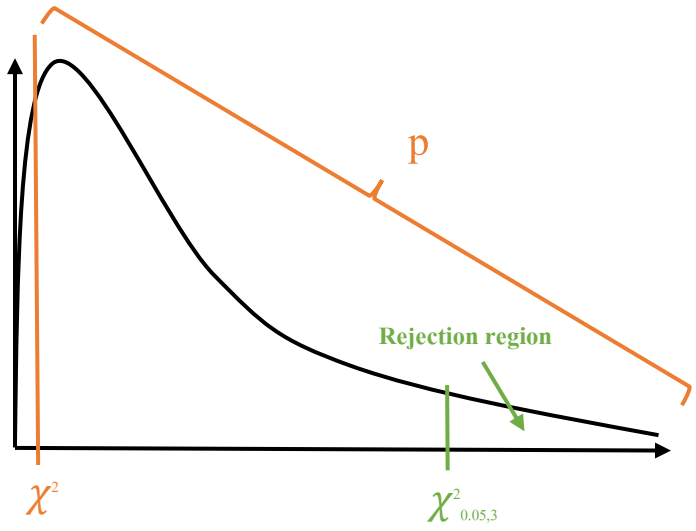
H_0 : ทั้งมอเตอร์ไซค์รับจ้างและรถประจำทางมีอัตราส่วนในการโดยสารในแต่ละช่วงราคาเท่ากัน

$p_{1j} = p_{2j}$ สำหรับ $j = 1, 2, 3, 4$

H_a : ทั้งมอเตอร์ไซค์รับจ้างและรถประจำทางมีอัตราส่วนในการโดยสารในแต่ละช่วงราคาไม่เท่ากัน

จะใช้ significance level อยู่ที่ 0.05

โดยคำนวณค่า Test statistic อยู่ที่ 0.447 และจะได้ค่า p-value อยู่ที่ 0.932 ค่า Degree of freedom เท่ากับ 3



ค่า p-value มีค่ามากกว่าค่า significance level ดังนั้นจึงไม่ปฏิเสธ Null hypothesis จึงสรุปได้ว่า ในการเดินทางระหว่างมอเตอร์ไซค์รับจ้าง และรถประจำทางมีอัตราส่วนในการโดยสารในแต่ละช่วงราคาเท่ากัน

* ค่าคำนวณต่างๆอยู่ในไฟล์ **“Categorical data analysis.xlsx”** และ **“assignment1.ipynb”**