Report on the Homogeneity of Drug related Crimes and Offender's Education level

JUNE 27

PRASANT KARUNAMURTHI
A19EC0148

UTM UNIVERSITI TEKNOLOGI MALAYSIA
SCHOOL OF COMPUTING
Faculty of Engineering

Table of Contents

Report on the Homogeneity of Drug related Crimes and Offender's Education level

1.0 Does Education level of an individual dictate their probability of committing a drug offence?

This report will focus on determining if there is a proper measurable relationship between someone's education and if it influences their overall probability on committing a drug related offence. A sample of 55,268 people's drug usage statistic were collected which. This dataset covers the usage of 13 drugs across 17 age groups. The data presented in this study is secondary and has been collected by their respected owner, the National Anti-Drug Agency. Before getting into the data, it is important to establish a hypothesis for the study. The hypothesis of this study is as follows.

The Education level of an individual does not dictate the probability of them committing a drug related offence.

1.1 Dataset

This report's dataset will focus on drug usage by age groups. The dataset is as follows.

STATISTICS OF DRUG ADDICTS BY CASE STSTUS, 2016 - 2018					
CATAGORY		PER CASE/ PER HEAD	2016	2017	2018
CASE STATUS/ YEAR	New Case	Per case	22,932	18,440	17,474
		Per Head	22,814	18,112	17,315
	Relapse Case	Per case	7,921	7,482	7,793
		Per Head	4,648	3,242	2,908
	Total	**Per case**	**30,844**	**25,922**	**25,267**
		Per Head	**27,462**	**21,354**	**20,223**

Source: https://www.adk.gov.my/en/public/drugs-statistics/

Table 1

Notes:

- **New Case** - Newly detected drug addicts to receive treatment/rehabilitation/ supervision with NADA
- **Relapse Cases** - Drug addicts who formerly detected and received treatment/ rehabilitation program with NADA
- 'Per Case' data refers to those who have one or more offences in the current year. Whilst the 'Per Head' data refers to only once detected within five (5) years period.

CATAGORY		PER CASE/PER HEAD	2016	2017	2018
AGE WHILE DETECTED/ YEAR	Adolescence (13-17 Year)	Per Case	826	694	550
		Per Head	810	683	541
	Youth (19-39 Year)	Per Case	22,709	18,823	18,417
		Per Head	20,335	15,812	15,131
	Adult (>=40 Year)	Per Case	7,309	6,405	6,300
		Per Head	6,317	4,859	4,551
	TOTAL	**Per Case**	**30,844**	**25,922**	**25,267**
		Per Head	**27,462**	**21,354**	**20,157**

Source: https://www.adk.gov.my/en/public/drugs-statistics/

Table 2

Notes:

- Per Case' data refers to those who have one or more offences in the current year. Whilst the 'Per Head' data refers to only once detected within six (6) years period.

CATAGORY		PER CASE/ PER HEAD	2016	2017	2018
ACADEMIC QUALIFICATION/ YEAR	Never Attended School/ Drop Out	Per Case	3,351	3,751	4,044
		Per Head	2,992	3,293	3,563
	Primary School	Per Case	2,898	2,370	2,118
		Per Head	2,529	1,848	1,588
	LCE/SRP/PMR	Per Case	11,730	9,082	8,451
		Per Head	10,297	7,192	6,459
	MCE/SPM/SPMV	Per Case	11,314	9,311	9,114
		Per Head	10,189	7,760	7,283
	HSC/STP/STPM	Per Case	218	166	166
		Per Head	205	135	132
	Diploma	Per Case	506	473	540
		Per Head	475	439	479
	Degree/Master/PhD	Per Case	96	113	114
		Per Head	92	106	103
	Others (E.G: Certificate of specific skills)	Per Case	731	656	720
		Per Head	683	581	616
	TOTAL	**Per Case**	**30,844**	**25,922**	**25,267**
		Per Head	**27,462**	**21,354**	**20,223**

Source: https://www.adk.gov.my/en/public/drugs-statistics/

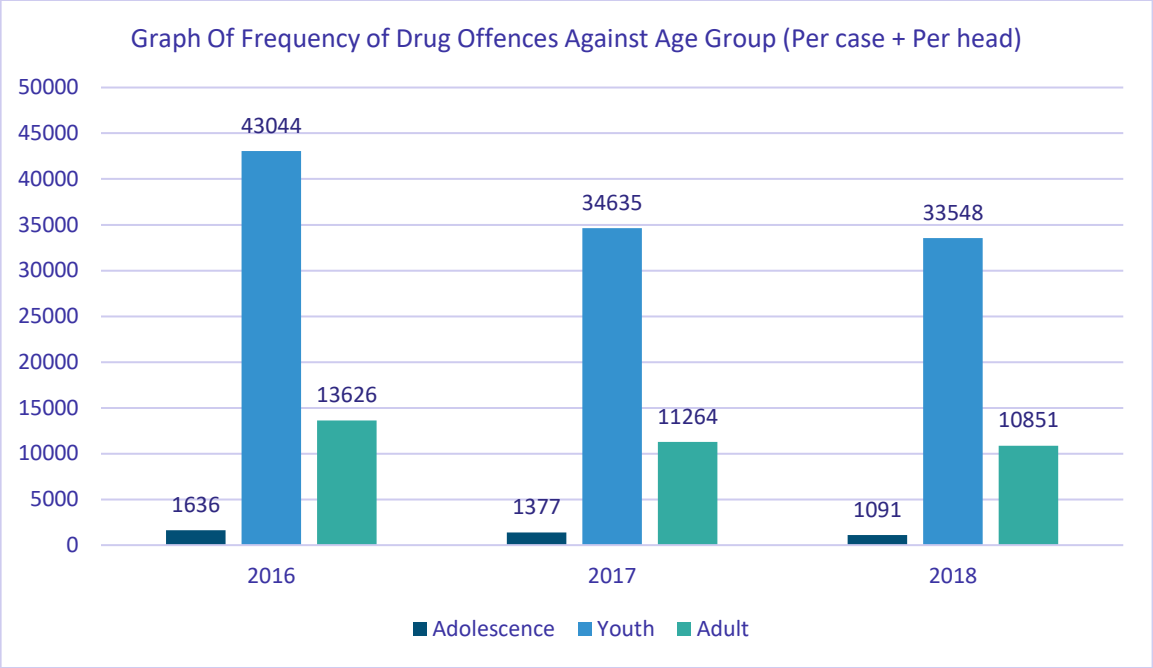
Table 3

1.2 Analyzing Raw Data

Table 1, 2, and 3 shows the raw data related for the study of Homogeneity of Drug related Crimes and Offender's Education level. To find if there exist a relationship between education level and possible drug offence, we first need to analyze the numbers on a timely manner; in this case throughout the year 2016, 2017 and 2018.

Looking at Table 1, it shows the number of drug offence throughout the year 2016, 2017 and 2018 on a per case and per head basis. Looking at the numbers, there is a clear decrease in number of drug offences both per case and per head. This seems to be a good indication that any type of drug related countermeasure is working as the numbers are going down. This however does not help with this study as the case/year statistic only acts as an averaging factor to get more consistent numbers.

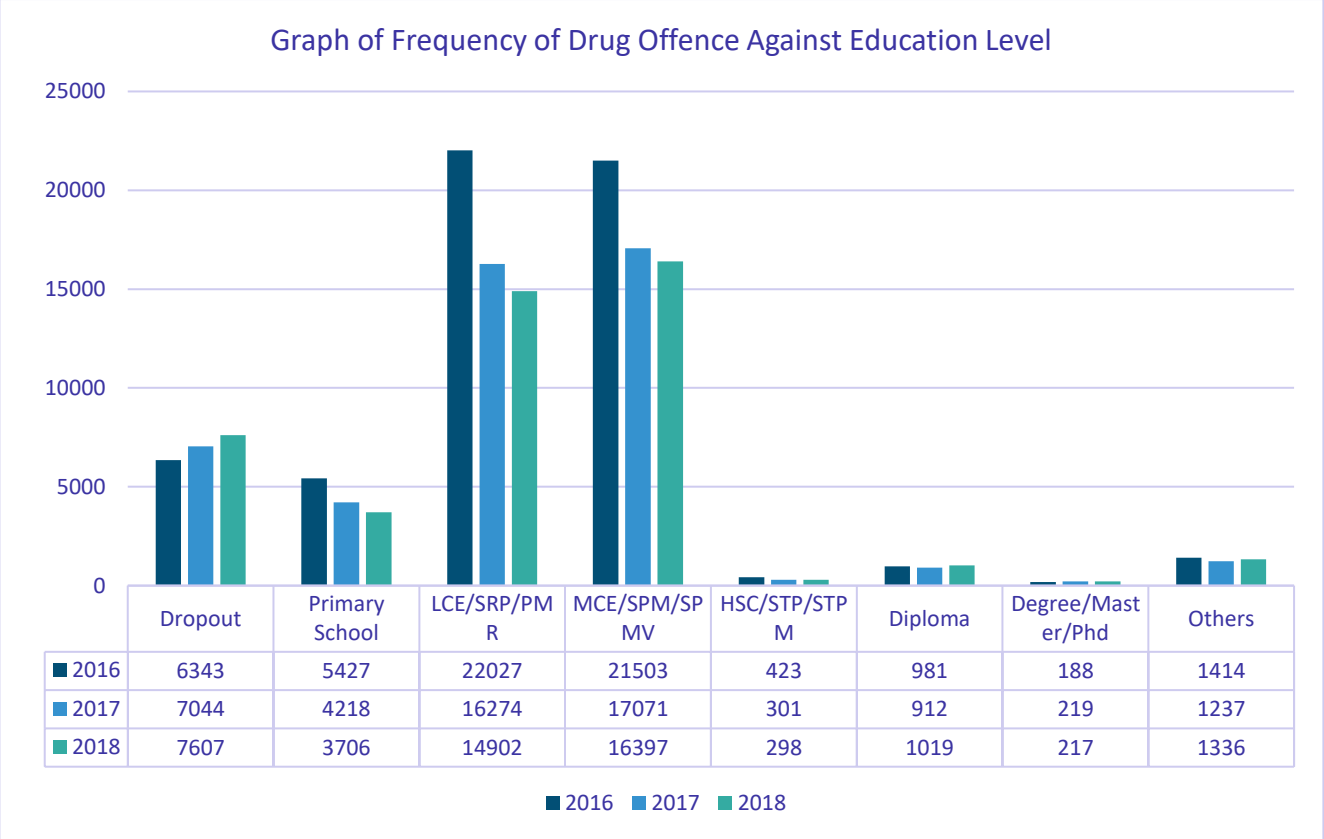
Moving on to Table 2, we can see the numbers when it comes to the Age of the drug offender. With this data, we can make a proper understanding on the number of drug offences and their age demographic. The data consists of 3 age demographic which are Adolescence (13-18 Year), Youth (19-39 Year), and Adult (>=40 Years) throughout the year 2016, 2017, and 2018.



Clustered Column 1

If we observe the clustered column, the youth age demographic has the highest numbers throughout the 3 years and although the numbers are gradually decreasing, it has substantially higher numbers than the other 2 age groups. To explain why this is the case, we need to look at the age range. Youth are made up of people from ages 19 to 39 years. This age period is admittedly where a person goes through a lot of changes someone's lifestyle changes. It is not surprising to see the highest number of drug related cases in this demographic. On the other end of the spectrum, Adolescence has the lowest cases, and this is not a small number especially considering the age group. But it is very miniscule when compared to the other numbers.

Table 3 shows the academic qualification and the correlating number of drug offences throughout the year 2016, 2017, and 2018. This is arguably the most important data as it directly relates to the study hypothesis. It gives us direct numbers to the question asked in the study which is: does academic qualification and the probability of someone committing a drug offence related?



	Dropout	Primary School	LCE/SRP/PMR	MCE/SPM/SPMV	HSC/STP/STPM	Diploma	Degree/Master/Phd	Others
2016	6343	5427	22027	21503	423	981	188	1414
2017	7044	4218	16274	17071	301	912	219	1237
2018	7607	3706	14902	16397	298	1019	217	1336

Clustered Column 2

Clustered Column 2 shows that the lower levels of education levels, e.g. dropouts, primary, and secondary education demographic yields the highest numbers of drug related cases. However, this data alone cannot conclude this study as there is no trend following the education level and number of cases. For example, Diploma has higher numbers than STPM even though it is a higher education level. Proper statistical analysis should be conducted with a proper null and test hypothesis to conclude the study.

The Table below shows the raw data for the sample individual by age group who has used said drug types. This data will be used to make a proper conclusion for this study and to answer the hypothesis. For simplicity, the ages from 12 to 21 will be grouped together.

age	n	alcohol-use	alcohol-frequency	marijuana-use	marijuana-frequency	cocaine-use	cocaine-frequency	crack-use	crack-frequency	heroin-use	heroin-frequency
12	2798	3.9	3	1.1	4	0.1	5	0	-	0.1	35.5
13	2757	8.5	6	3.4	15	0.1	1	0	3	0	-
14	2792	18.1	5	8.7	24	0.1	5.5	0	-	0.1	2
15	2956	29.2	6	14.5	25	0.5	4	0.1	9.5	0.2	1
16	3058	40.1	10	22.5	30	1	7	0	1	0.1	66.5
17	3038	49.3	13	28	36	2	5	0.1	21	0.1	64
18	2469	58.7	24	33.7	52	3.2	5	0.4	10	0.4	46
19	2223	64.6	36	33.4	60	4.1	5.5	0.5	2	0.5	180
20	2271	69.7	48	34	60	4.9	8	0.6	5	0.9	45
21	2354	83.2	52	33	52	4.8	5	0.5	17	0.6	30
22-23	4707	84.2	52	28.4	52	4.5	5	0.5	5	1.1	57.5
24-25	4591	83.1	52	24.9	60	4	6	0.5	6	0.7	88
26-29	2628	80.7	52	20.8	52	3.2	5	0.4	6	0.6	50
30-34	2864	77.5	52	16.4	72	2.1	8	0.5	15	0.4	66
35-49	7391	75	52	10.4	48	1.5	15	0.5	48	0.1	280
50-64	3923	67.2	52	7.3	52	0.9	36	0.4	62	0.1	41
65+	2448	49.3	52	1.2	36	0	-	0	-	0	120

Table 3

2.0 Statistical Calculative Data Analysis

2.1 2-Sample Test

Sample test question: Do younger people have higher tendencies to partake in drug related crimes repeatedly than older people in 2018?

	Drug related crimes (Per head)	Drug related crimes (Per case)
Younger people (19-39 Year)	18,417	15,131
Older people (>=40 Year)	6,300	4,551

	Group 1: Subjects are younger	Group 2: Subjects are older
Per head/ Per case	$X_1 = 18,417$	$X_2 = 4,551$
Subjects in group	$n_1 = 33,548$	$n_2 = 10,851$

$\hat{p}_1 = 18,417/33,548 = 0.5490$ $\hat{p}_2 = 4,551/10,851 = 0.4194$

"p-bar" = (18,17+4,551)/ (33,548+10,851) = 0.5173

"q-bar" = 1 – 0.5173 = 0.4827

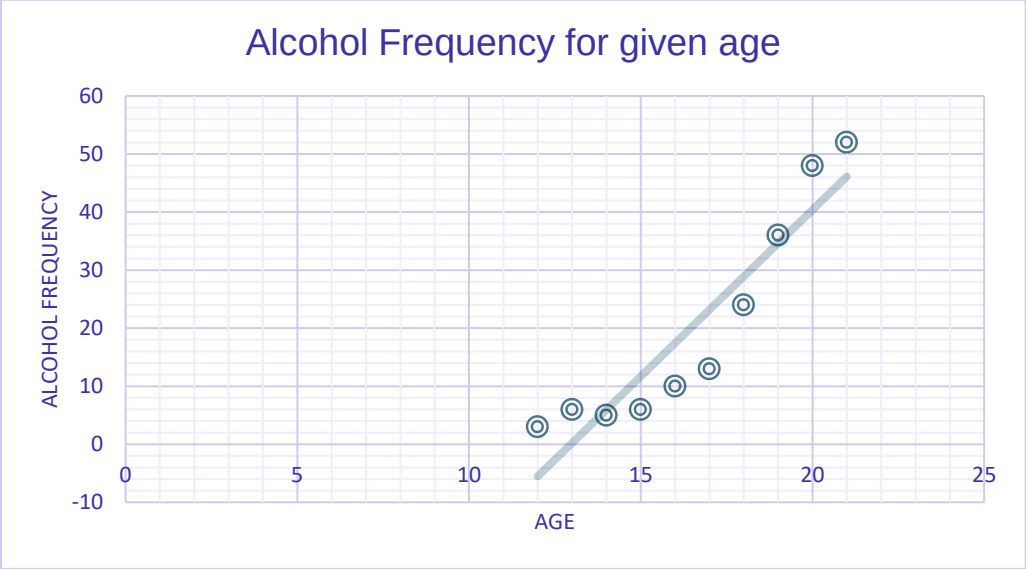
H_0: $P_1 = P_2$ Significance level: 0.05 Z = 23.4842

H_1: $P_1 > P_2$ $Z_{0.05} = 1.6450$

Since Z=23.4842 > 1.6450, reject H_0. There is no sufficient evidence to support the claim that younger people are more likely to make repeated drug offences on a significance level of 0.05.

2.2 Pearson Correlation Test

Correlation test is used to determine if there exists a relationship between two continuous variables. Correlation test also determines the strength of the relationship between these variables (i.e., how close the relationship is to being a perfectly straight line). For this study, the 2 continuous variables are age and alcohol frequency. This test is being carried out to determine if there exists a pattern in the consumption of alcohol among subjects from age 12 to 21 years. To understand the correlation better, a scatter plot would be useful.



Scatterplot 1

Alcohol is a very common type of substance being consumed by the younger population and from the scatterplot above, there is a positive correlation between the variable age and alcohol frequency. As age increases the alcohol frequency also seems to increase. To understand the strength of the correlation better, proper statistical analysis should be carried out to identify correlation coefficient.

Age, y	Alcohol Frequency, x	xy	Y^2	X^2
12	3	36	144	9
13	6	78	169	36
14	5	70	196	25
15	6	90	225	36
16	10	160	256	100
17	13	221	289	169
18	24	432	324	576
19	36	684	361	1296
20	48	960	400	2304
21	52	1092	441	2704
165	**203**	**3,823**	**2,805**	**7,255**

Correlation Coefficient, $r = [(3823 - (203x165)/10] / [[(7255) - (203)^2/10]x[(2,805)-(165)^2/10]]^{1/2}$

$r = 0.9312$

Correlation coefficient, r from the calculation is 0.9312 which indicates a relatively strong positive relationship between age and alcohol frequency. This conclusion reflects our understanding from the observation of scatterplot 1. From the correlation test, it can be concluded for the age group of 12 to 21 years, as the age increases the youth are more inclined to consume alcohol indicated by the increasing alcohol frequency.

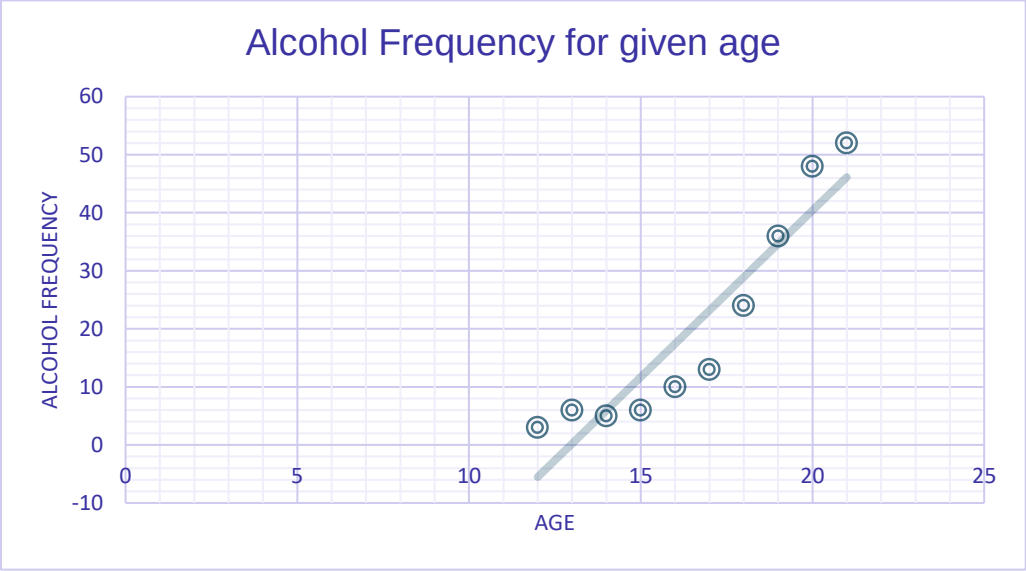
2.3 Regression Test

Regression test allows us to only use one independent variable and infer a relationship between said variable and a dependent variable. A simple linear regression test provides us with relationship between variables as a straight line. For our study, we can use the same variables as correlation test as the coefficient is higher than 0.5. for the variables, the age group(x) will act as independent variables whereas the alcohol frequency(y) will act as dependent variables.

Age, x	Alcohol Frequency, y	xy	x^2
12	3	36	144
13	6	78	169
14	5	70	196
15	6	90	225
16	10	160	256
17	13	221	289
18	24	432	324
19	36	684	361
20	48	960	400
21	52	1092	441
165	**203**	**3,823**	**2,805**

$b_1 = [(3823) – ((165)(203))/10)] / [(2805) – (165)^2/10] = 5.7394$

$b_0 = 20.3 – 5.7394(16.5) = -74.4001$



Scatterplot 1

"y-hat" = -74.4001 + 5.7394x

- **b_0** is the estimated average value of Y when the value of X is zero. For this test, no age group had 0 alcohol frequency, so b_0 = -74.4001 just indicates that, for frequencies within the range of age observed, -74.4 is the frequency not explained by the age.

- **b_1** measures the estimated change in the average value of Y as result of one-unit change in X. In this case, b_1 = 5.7394 tells us that the average alcohol frequency increases by 5.7394, on average, for each addition in age.

Age, x	Alcohol Frequency, y	"y-hat"	(y-hat – y-bar)^2	(Y_i – y-bar)^2
12	3	-5.5273	667.0494	299.29
13	6	0.2121	403.5237	204.49
14	5	5.9515	205.8794	234.09
15	6	11.6909	74.1166	204.49
16	10	17.4303	8.2351	106.09
17	13	23.1697	8.2351	53.29
18	24	28.9091	74.1166	13.69
19	36	34.6485	205.8794	246.49
20	48	40.3879	403.5237	767.29
21	52	46.1273	667.0494	1135.69
165	**203**	**203**	**2717.6084**	**3264.9**

y-bar = 203/10 = 20.3

SSR = total of (y-hat – y-bar)^2 = 2717.6084

SST = total of (Y_i – y-bar)^2 = 3264.9

R^2 = SSR/SST = 2717.6084/3264.9 = 0.8324

- 83% of the variation in alcohol frequency is explained by variation in age.

$S_\varepsilon = ((SSE)/n\text{-}k\text{-}1)^{1/2} = [(3264.9)/10\text{-}1\text{-}1]^{1/2} = 20.2018$

- Standard error of the regression slope coefficient (b_1) is estimated by S_{b1}.

  $S_{b1} = S_\varepsilon / [(\text{total of } x^2) - (\text{total of } x)^2/n]^{1/2} = 20.2018 / [(2805)-(165)^2/10]^{1/2} = 2.2241$

A relatively small sample standard error of the estimate allows us to infer that the variation of observed y values from the regression line is small and they are packed together. This is further proven by scatterplot 1.

- Inferring the slope
  - Is there a linear relationship between x and y?
- Null and alternate hypothesis
  - H0: β 1= 0 (no linear relationship)
  - H1: β 1$\neq$ 0 (linear relationship does exist)
- Test statistic:
  - t = (5.7394 – 0)/2.2241 = 2.58

d.f. = 10-2 = 8

a =.05

a/2=.025

$t_{\alpha/2}$ = 2.3060 (refed table)

test statistic t = 2.58 falls outside of critical region, thus H_0 is rejected. There is sufficient evidence to conclude that age has an influence on alcohol frequency.

2.4 Chi-Square Test

In the context of this study, chi-square test is used to identify if the drug related offences occur with equal proportion to the education level of the offenders. Data used in this test is from the year 2018.

Education level	Primary	PMR	SPM	STPM	Diploma	Degree/PhD
Drug offences	3,706	14,902	16,397	298	1,019	217

- Study claim: Drug offence happen with same proportion, thus $p_1 = p_2 = p_3 = p_4 = p_5 = p_6$.
- Test Hypothesis:
  - **H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6$**
  - **H_1: At least 1 of the 6 proportions is different from others.**
- Expected frequency, E = n/k = 36,539/6 = 6089.8

Education level	Primary	PMR	SPM	STPM	Diploma	Degree/PhD
Observed Drug offences (O)	3,706	14,902	16,397	298	1,019	217
Expected Drug offences (E)	6089.8	6089.8	6089.8	6089.8	6089.8	6089.8
$(O - E)^2/E$	933.1180	12,751.6288	17,445.2974	5,508.3824	4,222.3082	5,663.5324

- Critical value, X^2 = total of $(O - E)^2/E$ = 46,524.2672

- - K-1 = 6-1 = 5
  - a = 0.05
  - $X^2_{5,0.05}$ = 11.0705

Since critical value is greater than $X^2_{5,0.05}$, null hypothesis is rejected. We reject the claim that the drug related offences happen with equal proportion throughout the education levels. From this conclusion, we can conclude that there is no proportional relationship between a person's education and the probability of them committing a drug related offence.

3.0 Discussion

For this section of the report, we will be discussing the various results obtained from the studies carried out and ultimately try to piece together a proper conclusion to the initial hypothesis.

Firstly, the 2-sample test. This test was carried out to find out if younger or older people had the tendencies to commit continuous drug related offence. the age group were separated into younger (19-39) and older (>=40) demographic. To figure out the tendencies, the drug cases were separated into per head and per case. Per head data refers to those who have one or more offences in the current year. Whilst the 'Per Head' data refers to only once detected within five years period. This classification is used as it is easier to make a comparison for hypothesis testing. The 2-sample test concluded with a conclusion stating that there is no sufficient evidence to support the claim that younger people are more likely to make repeated drug offences. This conclusion further reinforces our initial hypothesis and proves that the idea of higher education can reduce drug crimes is not valid.

Next is the Pearson's Correlation test. Alcohol frequency and age was chosen to be the 2 variables for this test. Although alcohol is not an illicit drug, it is a substance which promotes change in an organism's physiology and psychology. In this test I wanted to determine how strongly related are the variables age and alcohol frequency. The correlation coefficient which was obtained was 0.9312 which indicates a relatively strong relationship between the 2 variables. What the test proves is that from the ages 12 to 21, as the person's age increases there is a high probability of alcohol frequency increasing hinting to a possible situation where younger demographic are more inclined to consume alcohol as they grow older.

Chi-square test are conducted to determine if there exists a proportion between variables. In the case of this study, education level and number of drug offences. I wanted to find out if the drug offences across different education levels had proportion which could indicate if they are proportional or not. The study concluded with a result stating that education level does not affect a person's probability of committing a drug related offence.

4.0 Conclusion & Reflection

To give a final verdict to this study is difficult especially considering the number of variables in play. But based on the various tests conducted, I believe that it is safe to assume that an individual no matter how high or low their education level is, the odds of them committing a drug related offence is not dictated by said education. This assumption is based on the data collected from the sample throughout 2016 to 2018 and is no means based on bias.

I also believe that no one factor leads to an individual committing a drug offence or reduces the probability for that matter. There are a lot of factors at play such as living environment, religious believes, health, peer pressure and much more. Everyone has their own struggles that they must go through, and it is not fair to assume the reasons of a person action without knowing them individually. With that said, it is high time to wrap up this study as a proper conclusion to they hypothesis has been made. Thank you for spending your precious time reading through this report.

To reflect on the project, it has provided me with new experience of analyzing data in a proper way and has taught me a lot about statistics. After completing this project, I have a new respect for the people working in statistical data analysis including the people teaching as it proves to be a vast field where data accuracy is at upmost importance. I am very thankful to my lecturer who has taught me and to the people who helped with this project.