UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SCSI 2143-08
PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2:
DEMOGRAPHY AND EPIDEMIOLOGY OF HIV AMONG ADOLESCENTS

LECTURER:
DR. SUHAILA MOHAMAD YUSUF

SUBMITTED BY:

CHOY WAN LING	**A18CS0049**
LEE CHOI WEI	**A18CS0095**
TANG KUAN YEW	**A18CS0261**
TOON SHU HUI	**A18CS0268**

Semester 2 2018/2019

TABLE OF CONTENTS

1.0 INTRODUCTION

In 2017, an estimated of 36.9 million people were living with HIV worldwide. Of these 36.9 million of people, approximately 3.0 million were children and adolescents (aged between 10 to 19). There are approximately 4900 people were newly infected with HIV every day. Meanwhile, there are approximately 2,580 people died every day from AIDS related causes due to inadequate access to HIV prevention, care and treatment services.

Hence, we decided to conduct a study on the Demography and Epidemiology of HIV among Adolescents in the world. The purpose of this study is to compare and analysis the difference between estimated number of male adolescents living with HIV over female adolescents, the proportion of adolescents living with HIV in different continents as well as the relationship between the estimated number of adolescents living with HIV and the number of AIDS deaths among adolescents.

We hope that we can find out the relationships among these variables through interpretation and analysis by conducting the Two-Sample Test (proportion), Chi-square Goodness of Fit Test, Chi-square Independent Test, correlations testing and.

2.0 METHODOLOGY

2.1 DATA COLLECTION

In this study, we have collected secondary data from UNICEF Data.
(source: https://data.unicef.org/resources/dataset/gender-and-hiv-data/)

The data consists of adolescent's population of each continent in the world, the estimated number of male and female adolescents living with HIV in 2016, and the number of AIDS deaths among adolescents in 2016. The data is well-organized in a statistical table.

2.2 DATA INTERPRETATION & PRESENTATION

We used R-Studio to do calculation on hypothesis testing and correlation. Besides, we also use R-Studio to make graphical presentation of our data and findings.

3.0 DATA ANALYSIS AND RESULTS

3.1 HYPOTHESIS TESTING (TWO-SAMPLE TEST)

Test:

To test whether there is sufficient evidence to say that **proportion of girl living with HIV in 2016 is different from proportion of boy living with HIV in 2016** with significance level of 0.05.

Calculation:

H_0: p1 = p2

[proportion of girl living with HIV in 2016 = proportion of boy living with HIV in 2016]

H_1: p1 ≠ p2

[proportion of girl living with HIV in 2016 ≠ proportion of boy living with HIV in 2016]

SUMMARY INDICATORS	Estimated number of adolescents (aged 10-19) living with HIV, 2016 Total	Girls	Boys
East Asia and the Pacific	60000	26000	34000
Eastern Europe and Central Asia	21000	14000	6800
Latin America and the Caribbean	77000	34000	43000
Middle East and North Africa	4000	2000	2000
South Asia	130000	64000	70000
Sub-Saharan Africaa/	1700000	1000000	730000
Eastern and Southern Africa	1300000	750000	540000
West and Central Africa	450000	260000	190000
World	**2100000**	**1200000**	**900000**

Test Statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(0.5714 - 0.4286) - (1200000 - 900000)}{\sqrt{(0.5 \times 0.5)\left(\frac{1}{1200000} + \frac{1}{900000}\right)}} = -614816753.1876$$

Z-critical value:

This is a two-tailed test, so ∝ should divide by 2 (0.05/2 = 0.025)

Z ∝= 0.025 = -1.96 and 1.96

Alternative Way:

By using R-studio,

```
> x1 = 1200000
> x2 = 900000
>
> n1 = 2100000
> n2 = 2100000
>
> phat1 = x1/n1
> phat2 = x2/n2
>
> pbar = (x1+x2)/(n1+n2)
> qbar = 1-pbar
>
> z = ((phat1-phat2)-(x1-x2))/sqrt(((pbar*qbar)/n1)+((pbar*qbar)/n2))
>
> alpha = 0.025
> z.alpha = qnorm(alpha)
```

Values	
alpha	0.025
n1	3742000
n2	3742000
pbar	0.50318011758418
phat1	0.574559059326563
phat2	0.431801175841796
qbar	0.49681988241582
x1	2150000
x2	1615800
z	-1461434789.03036
z.alpha	-1.95996398454005

Test statistic = -614816753.187554 (refer to z from R-studio printed screen above)

Z-Critical Value: Z_{α} = 0.025 = -1.95996398454005 (refer to z. alpha from R-studio printed screen above)

Conclusion:

Since the value of test statistic (-614816753.187554) is less than the Z-critical value (-1.95996398454005), so we reject H_0.

There is sufficient evidence at 0.05 significance level to support the claim that proportion of girl living with HIV in 2016 is different from the proportion of boy living with HIV in 2016.

3.2 CHI-SQUARE GOODNESS OF FIT TEST (ONE -WAY CONTINGENCY)

Test:

To test whether there is sufficient evidence to say that there is at least one of the proportion at continent is different from the others proportion at continent at 5% significance level.

Calculation:

H0: p1 = p2 = p3 = p4 = p5 = p6

H1: At least one of the proportions at continent is different from the others proportion at continent

Proportion of adolescents (aged 10 – 19) living with HIV at each continent in Year 2006

Continents	O	E	$(O-E)^2$	$\chi^2 = \frac{(O-E)^2}{E}$
East Asia and the Pacific	60,000	332000	73984000000	222843.3735
Eastern Europe and Central Asia	21,000	332000	96721000000	291328.3133
Latin America and the Caribbean	77,000	332000	65025000000	195858.4337
Middle East and North Africa	4,000	332000	107584000000	324048.1928
South Asia	130,000	332000	40804000000	122903.6145
Sub-Saharan Africa	1,700,000	332000	1871424000000	5636819.2770
TOTAL	1,992,000	1,992,000		6793801.2050

Since it is same probability for all the proportions, we use the total of the observed frequencies divide by 6 categories.

Degree of freedom = 6 categories – 1

= 5

Test statistic = 6793801.2050

Critical Value: $\chi^2_{5,0.05} = 11.071$ (refer to χ^2 table)

Alternative way:

We use R-Studio to get the chi-square, χ^2 directly

values	
alpha	0.05
exp	num [1:6] 222843 291328 195858 324048 122904 ...
expAcc	num [1:6] 332000 332000 332000 332000 332000 332000
expprob	332000
noAcc	num [1:6] 60000 21000 77000 4000 130000 1700000
x2	6793801.20481928
x2.alpha	11.0704976935164
x2.pvalue	0

```
> noAcc <- c(60000, 21000, 77000, 4000, 130000, 1700000)
> expprob <- sum(noAcc)/6
> expAcc <- c (expprob, expprob, expprob, expprob, expprob, expprob)
> exp <- ((noAcc-expAcc)^2)/expAcc
> x2 <- sum(exp)
```

Since it is same probability for all the proportions, we use the total of the observed frequencies divide by 6 categories using R-studio.

Degree of freedom = 6 categories – 1

= 5

Test statistic = 6793801.20481928 (refer to x2 from R-studio printed screen above)

```
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=5, lower.tail=FALSE)
> x2.pvalue <- pchisq(x2, df=5, lower.tail=FALSE)
```

Critical Value: $\chi^2_{5,0.05} = 11.0705$ (refer to x2.alpha from R-studio printed screen above)

Conclusion:

Since the value of test statistic (6793801.20481928) > χ^2 critical value (11.0705), it is in the critical region, hence we reject H0.

There is sufficient evidence at 0.05 significance level to prove that there is at least one of the proportion at continent is different from the others proportion at continent.

3.3 CHI - SQUARE INDEPENDENCE TEST (TWO-WAY CONTINGENCY)

Test:

To test whether there is sufficient evidence to say that there is a relationship between adolescents who are infected with HIV and adolescents who are death because of AIDS with significance level of 0.05.

Calculation:

H_0 = No relationship between adolescents who are infected with HIV and those who are death because of AIDS

H_1 = There is a relationship between adolescents who are infected with HIV and those who are death because of AIDS

	Gender						**Total**
	Female			Male			
	Observe Frequency	Expected Frequency	$\frac{(o-e)^2}{e}$	Observe Frequency	Expected Frequency	$\frac{(o-e)^2}{e}$	
HIV	1,200,000	1,194,709.977	23.4235	900,000	905,290.0232	30.9120	**2,100,000**
AIDS	26,000	31,290.0232	894.3536	29,000	23,709.9768	1180.2772	**55,000**
Total	**1,226,000**			**929,000**			**2,155,000**

```
> female <- c(1200000,26000)
> male <- c(900000,29000)
> d <- data.frame(female, male)
> chisq.test(d, correct = FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 2129, df = 1, p-value < 2.2e-16
```

Test statistic = 23.4235 + 894.3536 + 30.9120 + 1180.2772

= 2128.9663 (refer to x-squared from R-studio printed screen above)

```
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df = 1, lower.tail = FALSE)
```

x2.alpha	3.84145882069413

Critical Value: $\chi^2_{1,0.05} = 3.841$ (refer to x2.alpha from R-studio printed screen above)

Conclusion:

Since the value of test statistics (2128.9663) > critical value (3.841), we can reject H_0.

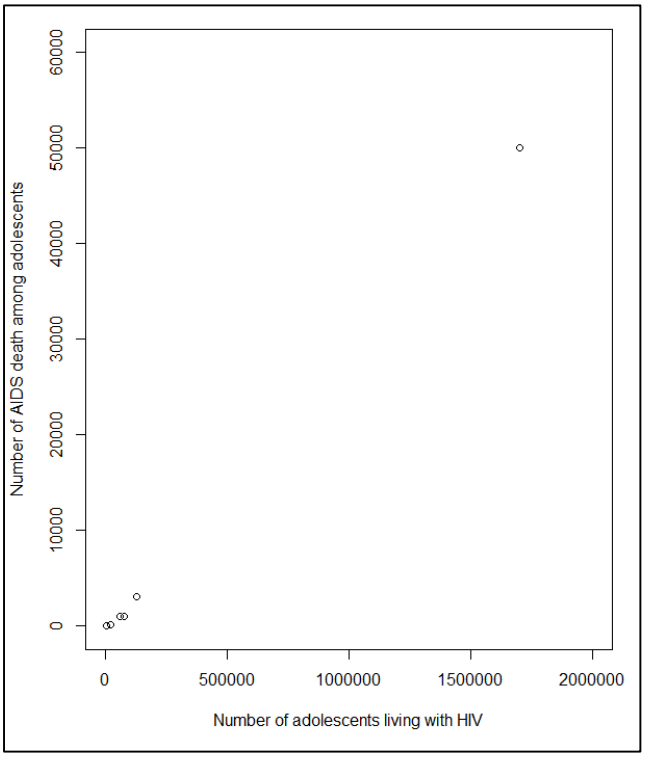
There is sufficient evidence at 0.05 significance level to prove that there is a relationship between adolescents who are infected with HIV and adolescents who are death because of AIDS.

3.4 CORRELATION TESTING

Test:

To find out the relationship between the number of adolescents living with HIV and the number of AIDS deaths among adolescents.

Calculation:



From the scatter plot above, we can clearly know that it is a linear relationship between the two variables. The data is ratio data. Hence, we apply Pearson's product-moment to get the correlation coefficient, r.

x	y	xy	x^2	y^2
60000	1000	60000000	3600000000	1000000
21000	200	4200000	441000000	40000
77000	1000	77000000	5929000000	1000000
4000	100	400000	16000000	10000
130000	3100	403000000	16900000000	9610000
1700000	50000	85000000000	2890000000000	2500000000
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$
1992000	55400	85544600000	2916886000000	2511660000

x: number of adolescents living with HIV

y: number of AIDS deaths among adolescents

The value of $\sum x, \sum y, \sum x^2, \sum y^2$ is then substituted into following equation,

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{\left[\left(\sum x^2\right) - \left(\sum x\right)^2/n\right]\left[\left(\sum y^2\right) - \left(\sum y\right)^2/n\right]}}$$

$$r = \frac{85544600000 - (1992000 \times 55400) / 6}{\sqrt{(2916886000000 - 1992000^2 / 6)\,(2511660000 - 55400^2 / 6)}}$$

$$r = 0.9998$$

Alternatively, we can use R-Studio to get the correlation coefficient, r directly.

```
> hiv <- c(60000, 21000, 77000, 4000, 130000, 1700000)
> aids <- c(1000, 200, 1000, 100, 3100, 50000)
>
> cor(hiv, aids)
[1] 0.999776
>
```

Conclusion:

From our analysis, the correlation coefficient, r = 0.9998. Hence, this indicates that there is a positive correlation between the number of adolescents living with HIV and the number of AIDS deaths among adolescents. It shows a strong positive correlation because r falls within +0.8 to +1.

3.5 Regression Testing

Test:

To examine the relationship between number of adolescents infected with HIV and population of adolescents.

Dependent variables (y): Number of adolescents infected with HIV.

Independent variables (x): Population of adolescents.

Calculation:

NUMBER OF POPULATION OF ADOLESCENTS, x	NUMBER OF ADOLESCENTS INFECTED WITH HIV, y	xy	x^2 x^2
300,190,000	60,000	18,011,400,000,000	90,114,036,100,000,000
51,586,000	21,000	1,083,306,000,000	2,661,115,396,000,000
110,572,000	77,000	8,514,044,000,000	12,226,167,184,000,000
73,653,000	4,000	294,612,000,000	5,424,764,409,000,000
340,270,000	130,000	44,235,100,000,000	115,783,672,900,000,000
232,263,000	1,700,000	394,847,100,000,000	53,946,101,169,000,000
1,108,534,000	**1,992,000**	**466,985,562,000,000**	**280,155,857,158,000,000**

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_1 = \frac{466985562000000 - \frac{1108534000 \times 1992000}{6}}{280155857158000000 - \frac{1108534000^2}{6}}$$

$$= 0.001313271$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_0 = \frac{1992000}{6} - 0.001313271 \times \frac{1108534000}{6}$$

$$= 89365.64988$$

From the calculation above, the estimated equation is $\hat{y} = 89365.6499 + 0.0013x$.

T – test: Is there a relationship between x and y?

$\hat{y} = 89365.6499 + 0.0013x$

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

NUMBER OF POPULATION OF ADOLESCENTS, x	NUMBER OF ADOLESCENTS INFECTED WITH HIV, y	$\hat{y}$	$(y_i - \hat{y})^2$
300,190,000	60,000	483,597	179,434,095,821
51,586,000	21,000	157,112	18,526,496,442
110,572,000	77,000	234,577	24,830,418,070
73,653,000	4,000	186,092	33,157,509,254
340,270,000	130,000	536,233	165,024,877,082
232,263,000	1,700,000	394,390	1,704,617,402,763
1,108,534,000	**1,992,000**	**1,992,000**	**2,125,590,799,433**

$$s_\varepsilon = \sqrt{\frac{2125590799433}{6-1-1}}$$

$$= 728970.3011$$

$$S_{b_1} = \frac{728970.3011}{\sqrt{280,155,857,158,000,000 - \frac{1,108,534,000^2}{6}}}$$

$$= 0.002656$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$t = \frac{0.0013 - 0}{0.002656}$$

$$= 0.4895$$

Degree of freedom $= 6 - 2 = 4$

$\alpha = 0.05 / 2 = 0.025$

$t_{\alpha/2,4} = 2.776$

Decision: $t < t_{\alpha/2,4}$, Fail to reject H_0.

Conclusion:

There is insufficient evidence at 0.05 significance level to say that there is a relationship between number of adolescents infected with HIV and population of adolescents.

4.0 DISCUSSION AND CONCLUSION

From the first hypothesis testing (two-sample test), we can conclude that proportion of girl living with HIV in 2016 is different to the proportion of boy living with HIV in 2016, which means there is a difference in between each gender of adolescents of the world living with HIV in 2016. Similarly, we can conclude that the proportion of adolescents living with HIV in each continent of the world is also different using chi-square goodness of fit test.

Next, we use chi-square independence test to prove that there is a relationship between adolescents who are infected with HIV and adolescents who are death because of AIDS. This claim is supported by correlation testing which proves that there is a strong positive correlation between the number of adolescents living with HIV and the number of AIDS death among adolescent.

Finally, from the regression testing, we do not have sufficient evidence to claim that there is a relationship between the number of adolescents infected with HIV and the world population of adolescents.

In short, we can tell that the number of adolescents living with HIV in 2016 differs from gender and the continent they are living in. Meanwhile, we could not claim that the population of adolescents do affect the number of adolescents infected with HIV. However, the strong relationship in between adolescents who are infected with HIV and adolescents who are death because of AIDS shows that HIV affects in AIDS death.

As a conclusion, the conducted tests have shown relationship in between each variable. However, we have found out the weakness of our study, where we use a very large size of sample, causing all the values and frequencies to be extremely large. It may cause confusion and mistakes during calculation. Ways to prevent such mistakes including to check the values of each calculation carefully or to choose a dataset with smaller sample size.

5.0 REFERENCES

	Demographics			Epidemiology										
	Population (thousands), 2016		Adolescents as a % of total population, 2016	Estimated number of adolescents (aged 10-19) living with HIV, 2016			Adolescents living with HIV as a % of total HIV population, 2016	Number of new HIV infections among adolescents (aged 15-19), 2016			Number of AIDS deaths among adolescents (aged 10-19), 2016		
SUMMARY INDICATORS	Total	Age 10-19		Total	Girls	Boys		Total	Girls	Boys	Total	Girls	Boys
East Asia and the Pacific	2,288,860	300,190	13	60,000	26,000	34,000	2	15,000	5,800	9,400	<1,000	<500	<500
Eastern Europe and Central Asia	423,392	51,586	12	21,000	14,000	6,800	1	7,200	5,100	2,200	<200	<200	<100
Latin America and the Caribbean	632,043	110,572	17	77,000	34,000	43,000	4	19,000	7,900	11,000	<1,000	<500	<1,000
Middle East and North Africa	427,528	73,653	17	4,000	2,000	2,000	3	1,100	<1000	<1000	<100	<100	<100
South Asia	1,743,948	340,270	20	130,000	64,000	70,000	6	18,000	8,300	9,300	3,100	1,400	1,700
Sub-Saharan Africaa/	1,007,878	232,263	23	1,700,000	1,000,000	730,000	7	190,000	140,000	47,000	50,000	24,000	26,000
Eastern and Southern Africa	529,385	122,857	23	1,300,000	750,000	540,000	7	130,000	98,000	28,000	34,000	16,000	18,000
West and Central Africa	478,493	109,406	23	450,000	260,000	190,000	7	62,000	43,000	19,000	16,000	7,700	8,500
World	7,383,009	1,206,046	16	2,100,000	1,200,000	900,000	6	260,000	170,000	86,000	55,000	26,000	29,000

1. Unicef Data (2017): Demography and Epidemiology of HIV among Adolescents. Retrieved from: https://data.unicef.org/resources/dataset/gender-and-hiv-data/ ; Last surveyed on 26th April 2019.
2. Hiroshi Nishiura (2019): Estimating the incidence and diagnosed proportion of HIV infections in Japan: a statistical modeling study. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6338104/ ; Last surveyed on 4th May 2019.
3. Lecturer's note: Chapter 4 Probability, Random Variables and Probability Distributions
4. Lecturer's note: Chapter 5 (Part 1) Point Estimator
5. Lecturer's note: Chapter 5 (Part 2) Hypothesis Testing 1 Sample
6. Lecturer's note: Chapter 5 (Part 3) Hypothesis Testing 2 Samples
7. Lecturer's note: Chapter 6 Chi Square Test Contingency
8. Lecturer's note: Chapter 7 (Part 1) Correlation Analysis
9. Lecturer's note: Chapter 7 (Part 2) Linear Regression Model