SCSI 2143 PROBABILITY AND STATISTICAL DATA ANALYSIS

Project II – Inferential Statistics

Suicide Rates Report

Osama Abdelhameed Shehata Abdelnasser

Section 07

School of Computing, Faculty of Engineering

Universiti Teknologi Malaysia

Abstract- *This case study is about the statistical analysis of the relationship between suicide numbers per 100K for each group of people divided related to their age groups and the annual GDP per capita of participant countries.*

I. Introduction

Suicide rate overview 1985 to 2016 is secondary data collected by the United Nation Development Program, World Bank, and the World Health Organization to compare socio-economic information with suicide rates by year and country.

Helping countries to solve their socio-economic problems is an expert's responsibility. I choose comparison between Finland as a rich country and Russia as a poor country to show if there is a relationship between wealth and suicide rates with another comparison for Russia only as a growing country from poverty to show if growth could affect suicide's rates to be my research-based paper. Groups of people will be chosen in an appropriate way which will avoid elder people above 75 years old because they considered as a minority in different communities and below 15 years old due to the low impact of wealth on them.

II. Methodology

I will use R studio to assist me to process statistical analyses on the data-sets. This data set is analyzed through several statistical tests such as Hypothesis testing two samples, Correlation, Regression, ANOVA, and Chi-square tests.

Grouped data will provide different division and measurement ways like the comparison between Russia in poverty and wealth will measure by Hypothesis testing two, ANOVA test and Chi-square. However, the comparison between Finland as a rich country and Russia as a poor country in suicide rates will be done by Correlation and Regression.

III. Result and Discussion

1- Hypothesis Testing 2-Sample

Testing two dependent samples to measure if suicide number per 100K effected by the country's economy, so choosing Russia as improved from poverty during 1990 & 1991 and rich economy during 2010 & 2011, means after 20 years later. The mean GDP per capita between 1990 & 1991 is 3781$, however, the mean of 2010 & 2011 is 13266.5$ which means that GDP per capita after 20 years is greater by 251%. Ignoring gender and age impact by processing each element with its same gender and interval age class after 20 years.

μ1 is the mean of suicide number / 100K of Russia from 1990 to 1991.

μ1 is the mean of suicide number / 100K of Russia from 2010 to 2011.

$$H_0: \mu_D = 0$$
$$H_1: \mu_D > 0$$

```
#hypothesis testing two samples
Russia.before = c(28.57,55.04,67.68,66.83,6.64,7.46,12.5,19.72,29.51,56.76,67.86,66.36,6.62,7.46,12.15,18.54)
Russia.after = c(37.3,57.1,52.08,46.34,7.45,7.78,7.67,8,35.44,51.83,47.81,42.91,7.61,7.24,6.97,7.56)
D = Russia.before - Russia.after
D2 = D^2
#D2 is D square
sum.D = sum(D)
sum.D2 = sum(D2)
D.avr = sum.D/16
d = sum(D)^2/16
S.D = sqrt((sum.D2-d)/15)
t.dpendentest = D.avr/(S.D/sqrt(16))
t.criticalvalue = qt(1-0.025, 15)
```

d	607.74575625
D	num [1:16] -8.73 -2.06 15.6 20.49 -0.81 ...
D.avr	6.163125
D2	num [1:16] 76.213 4.244 243.36 419.84 0.656 ...
Russia.after	num [1:16] 37.3 57.1 52.08 46.34 7.45 ...
Russia.before	num [1:16] 28.57 55.04 67.68 66.83 6.64 ...
S.D	9.8561410425176
sum.D	98.61
sum.D2	2064.8985
t.criticalvalue	2.13144954555978
t.dpendentest	2.50123246954904

According to previous figures t critical value (t_{cv}) = 2.1315 at α = 0.025 with df = 16 - 1= 15 so to reject the null hypothesis t-test statistic should be more than 2.1315 so it rejected because test statistic is t = 2.5012, to conclude there is sufficient evidence to support the claim that the mean of suicide number per 100K of Russia from 1990 to 1991 is greater than the mean of suicide number per 100K of Russia from 2010 to 2011 among people from different age groups and genders.

Correlation

year	Russian GPA per capita	Russian suicide/100K	Finnish GPA per capita	Finnish suicide/100K
2005.00	5611.00	308.26	41202.00	172.72
2006.00	7313.00	286.09	43487.00	190.89
2007.00	9643.00	277.04	51089.00	174.49
2008.00	12359.00	256.40	56521.00	182.44
2009.00	9118.00	252.95	49878.00	182.68
2010.00	11307.00	223.72	48939.00	204.24
2011.00	15226.00	207.37	53809.00	159.42
2012.00	16413.00	197.83	50232.00	154.61
2013.00	17052.00	190.03	52572.00	157.23
2014.00	15319.00	173.28	52832.00	183.06

Measured data changes to fit the Correlation test to be GPA per capita with the sum of suicide numbers per 100K of different genders and group ages every year from 2005 to 2014, so the data changed as the table shows up.

```
> Russia.GDPcapita = c(5611, 7313,9643,12359, 9118, 11307, 15226, 16413, 17052, 15319)
> Russia.suicide = c(308.26,286.09,277.04,256.4,252.95,223.72,207.37,197.83,190.03,173.28)
> r1 = cor(Russia.suicide,Russia.GDPcapita)
> r1
[1] -0.9345643
> n=10
> tR = r1/ (sqrt((1-r1^2)/ (n-2)))
> tR
[1] -7.429427
```

H_0: There is no linear correlation between GDP per capita and suicide numbers per 100K in Russia.
H_1: The linear correlation exists between GDP per capita and suicide numbers per 100K in Russia.

$$\boldsymbol{\rho = 0}$$
$$\boldsymbol{\rho \neq 0}$$

Referring to the calculated correlation value for Russian GDP per capita to Russian suicide numbers per 100K for ten years respectively, the correlation value equals -0.9346 which means that there is a strong inversely relationship between GDP per capita and suicide numbers per 100K in Russia where considered as poor country improved by time.

```
> # Finnish correlation
> Finland.GDPcapita = c(41202, 43487,51089,56521, 49878, 48939,53809, 50232, 52572, 52832)
> Finland.suicide = c(172.72,190.89,174.49,182.44,182.68,204.24,159.42,154.61,157.23,183.06)
> r2= cor(Finland.suicide,Finland.GDPcapita)
> r2
[1] -0.2346364
> n=10
> tF = r2/ (sqrt((1-r2^2)/ (n-2)))
> tF
[1] -0.6827109
> df = n-2
> alfa = 0.05
> t.cv = qt(alfa/2,df)
> t.cv
[1] -2.306004
```

H_0: There is no linear correlation between GDP per capita and suicide numbers per 100K in Finland.
H_1: The linear correlation exists between GDP per capita and suicide numbers per 100K in Finland.

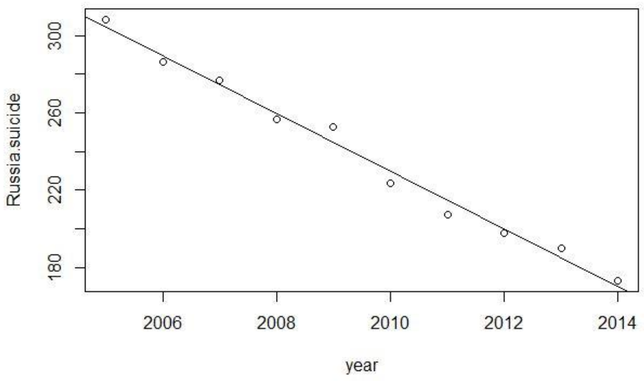
$$\boldsymbol{\rho = 0}$$
$$\boldsymbol{\rho \neq 0}$$

Whereas, based on the computed correlation value for Finnish GDP per capita to Finnish suicide numbers per 100K for ten years respectively, the correlation value equals -0.2346 which means that there is a weak inverse relationship between GDP per capita and suicide numbers per 100K in Finland where considered as rich country fluctuating at the same point, so at t critical value it shows that Russia had linear correlation exist at 0.05 alpha and Finland did not had.

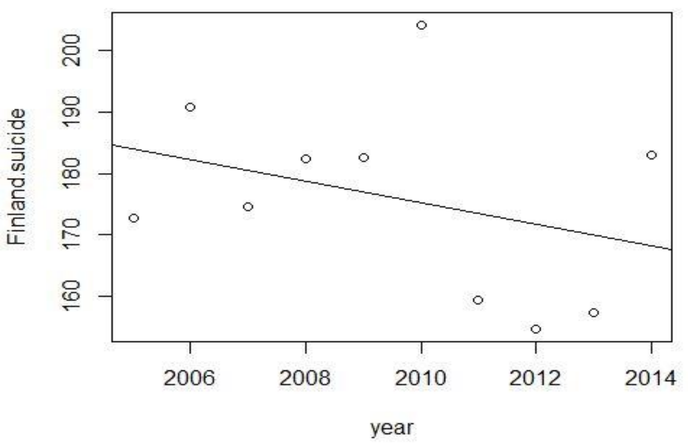
Regression

The dependent variable (y) in this case is the suicide numbers per 100K in each country whereas, the independent variable (x) is the respective ten years from 2005 to 2014. This analysis aims to test the existence of a linear relationship between the variable x and y for both countries Russia and Finland.

H_0: β1 = 0
H_1: β1 ≠ 0



```
> Russia.suicide = c(308.26,286.09,277.04,256.4,252.95,223.72,207.37,197.83,190.03,17
> year = c(2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014)
> cor(Russia.suicide,year)
[1] -0.9934233
> model = lm(Russia.suicide~year)
> model

Call:
lm(formula = Russia.suicide ~ year)

Coefficients:
(Intercept)         year
   30192.37       -14.91

> summary(model)

Call:
lm(formula = Russia.suicide ~ year)

Residuals:
   Min     1Q Median     3Q    Max
-7.567 -3.350  0.138  3.678  8.200

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 30192.3655  1220.6631   24.73 7.63e-09 ***
year          -14.9067     0.6074  -24.54 8.12e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.517 on 8 degrees of freedom
Multiple R-squared:  0.9869,    Adjusted R-squared:  0.9853
F-statistic: 602.2 on 1 and 8 DF,  p-value: 8.121e-09
```




```
> Finland.suicide = c(172.72,190.89,174.49,182.44,182.68,204.24,159.42,154.61,157.23,183.06)
> year = c(2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014)
> cor(Finland.suicide,year)
[1] -0.3357286
> model = lm(Finland.suicide~year)
> model

Call:
lm(formula = Finland.suicide ~ year)

Coefficients:
(Intercept)         year
   3701.450       -1.754

> summary(model)

Call:
lm(formula = Finland.suicide ~ year)

Residuals:
    Min      1Q  Median      3Q     Max
-17.182 -12.444  -1.222   7.835  28.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3701.450   3496.968   1.058    0.321
year          -1.754      1.740  -1.008    0.343

Residual standard error: 15.81 on 8 degrees of freedom
Multiple R-squared:  0.1127,    Adjusted R-squared:  0.001803
F-statistic: 1.016 on 1 and 8 DF,  p-value: 0.3429
```

According to the graphs generated of suicide numbers per 100K, the regression models are negative linear model which have straight-linear relationships. Through regression analysis, we can predict that number of suicide numbers per 100K would gradually decrease if GDP per capita increases. Russia will decrease at a faster rate than Finland.

$$y_{Russia} = 30192.37 - 14.91x$$

Russia's Y equation shows that the estimated sum changes in Russian suicide numbers per 100K decrease by 14.91. There is no suicide number per 100K equal to 0 during selected years, so 30192.37 indicates that for suicide numbers per 100K within the years observed is the portion of suicide number per 100K throughout the years in Russia.

$$R_{Russia}{}^{2} = \frac{SSR}{SST} = \frac{18772.06}{19884.52} = 0.944$$

94.4% of the variation in suicide numbers per 100K is explained by the variation in years.

$$y_{Finland} = 3701.45 - 1.754x$$

inland's Y equation shows that the estimated sum changes ın Finnish suicide numbers per 100K decrease by 1.754. There is no suicide number per 100K equal to 0 during elected years, so 3701.45 indicates that for suicide numbers ier 100K within the years observed is the portion of suicide umber per 1100K throughout the years in Finland.

$$R_{Finland}{}^{2} = \frac{SSR}{SST} = \frac{260.688}{2253.11} = 0.1157$$

11.57% of the variation in suicide numbers per 100K is explained by the variation in years.

In addition, correlation values between dependent and independent values are -0.9934 and -0.3357 for Russia and Finland respectively. To conclude, calculated data illustrates that Russia is closer to the mean throughout the years compared to Finland, also Russia has a stronger relationship than Finland.

Independence Chi Square Test

Returning back to improving poor country (Russia) in hypothesis testing two samples' data to show if suicide numbers per 100K have a relationship with years, gender and age-grouped combined or not.

	Russia before	Russia after	Total
90&10 Male 15-24	28.57	37.30	65.87
90&10 Male 25-34	55.04	57.10	112.14
90&10 Male 35-54	67.68	52.08	119.76
90&10 Male 55-74	66.83	46.34	113.17
90&10 Female 15-24	6.64	7.45	14.09
90&10 Female 25-34	7.46	7.78	15.24
90&10 Female 35-54	12.50	7.67	20.17
90&10 Female 55-74	19.72	8.00	27.72
91&11 Male 15-24	29.51	35.44	64.95
91&11 Male 25-34	56.76	51.83	108.59
91&11 Male 35-54	67.86	47.81	115.67
91&11 Male 55-74	66.36	42.91	109.27
91&11 Female 15-24	6.62	7.61	14.23
91&11 Female 25-34	7.46	7.24	14.70
91&11 Female 35-54	12.15	6.97	19.12
91&11 Female 55-74	18.54	7.56	26.10

The table explains the data structure by each rows' name for example the first one (90&10 Male 15-24) 90&10 refers to 1990 for Russian before data means before 20 years from 2010 which is the extend of 10, male refers to the gender and 15-24 refers to selected age-grouped.

H_0: No relationship between suicide numbers per 100K and years, gender & age-grouped combined in Russia.
H_1: Relationship exist between suicide numbers per 100K and years, gender & age-grouped combined in Russia.

```
> t = data.frame(Russia.before, Russia.after)
> x =chisq.test(t, correct = FALSE)
> x

        Pearson's Chi-squared test

data:  t
X-squared = 18.544, df = 15, p-value = 0.2351

> alpha <- 0.1
> x2.alpha <- qchisq(alpha,df=15,lower.tail=FALSE)
> x2.alpha
[1] 22.30713
```

Based on the result, it says that the chi-square test statistic value which is called x-squared value is equal to 18.544 while the critical value is equal to about 22.31 when alpha = 0.1 and df = 16 -1 = 15.

To conclude it shows that it failed to reject because the critical value is higher than chi-square test statistic value which means no relationship between suicide numbers per 100K and years, gender & age-grouped combined in Russia before & after.

ANOVA

Continuing with same data in chi-square table to measure whether the mean of Russia's suicide numbers per 100K during 1990-1991 and 2010-2011 periods, to explain is there are any difference between their means.

$_0$ represent the mean of Russia's suicide numbers per 100K uring 1990-1991
$_1$ represent the mean of Russia's suicide numbers per 100K uring 2010-2011

H_0: $\mu_0 = \mu_1$
H_1: at least one is different

```
> meanB = mean(Russia.before)
> meanB
[1] 33.10625
> meanA= mean(Russia.after)
> meanA
[1] 26.94312
> standard.before = sqrt(var(Russia.before))
> standard.before
[1] 25.42042
> standard.after = sqrt(var(Russia.after))
> standard.after
[1] 20.68952
> n = 16
> k = 2
> c = c(meanB,meanA)
> mean.all = mean(c)
> mean.all
[1] 30.02469
> standard.all = sqrt(var(c))
> standard.all
[1] 4.357987
> s = c(var(Russia.before), var(Russia.after))
> mean.standards = mean(s)
> mean.standards
[1] 537.1268
> F = (n * var(c))/mean.standards
> F
[1] 0.5657377
> df1 = k - 1
> df2 = k * (n-1)
> f.alpha = qf(0.9, df1, df2)
> f.alpha
[1] 2.880695
```

Based on the result, it says that F test statistic value is equal to 0.5657, while the critical value is equal to about 2.881 when alpha = 0.1, the numerator (df) = 2 -1= 1 and denominator(df) = 2 (16 - 1) = 30.
So, it illustrates that H is failed to reject because the critical value is higher than the test statistic value, which supports the claim the first mean that represents the mean of Russia's suicide numbers per 100K during 1990-1991 and the second mean that represent the mean of Russia's suicide numbers per 100K during 2010-2011 are considered equal at 0.1 alpha.

IV. Conclusion

In summary, hypothesis testing two samples explained that the mean of suicide number per 100K of Russia from 1990 to 1991 is greater than the mean of suicide number per 100K of Russia from 2010 to 2011, chi-square test shows no relationship between suicide numbers per 100K and years, gender & age-grouped combined in Russia and ANOVA illustrates the mean of Russia's suicide numbers per 100K during 1990-1991 is equal to the mean of Russia's suicide numbers per 100K during 2010-2011. Overall, the poor country's self-suicide rates will decrease when its economy getting better.

In the end, correlation test shows the strong inversely relationship between GDP per capita and suicide numbers per 100K in Russia where considered as poor country improved by time, the weak inverse relationship between GDP per capita and suicide numbers per 100K in Finland where considered as rich country fluctuating at the same point. Regression supports the same point when year data became the independent variable of Russian suicide numbers per 100K and Finnish suicide numbers per 100K by showing the high explained variation of Russian suicide numbers per 100K comparing to Finland's.