FACULTY ENGINEERING

SCHOOL OF COMPUTING

SECI 2143 SSECTION 10 – PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

TOPIC: CANCER CASES AND DEATH

LECTURER'S NAME: DR AZURAH BINTI ASMAH

NAME: CHONG HONG LEI

MATRIC NUMBER: A19EC0035

CONTENT

1.0 INTRODUCTION

"Cancer is the second leading cause of death globally and is responsible for an estimated 9.6 million deaths in the year of 2018. Globally, about 1 in 6 deaths is due to cancer." reported by world health organization (WHO). **Cancer -** is a generic term for a large group of diseases that can affect any part of our body. It is a genetic disease that caused by the changes of genes that control the way our cells function, especially how they grow and divide. Genetic changes that cause cancer can be inherited from our parents. They may also arise during a person's lifetime as a result of errors that occur as cell divide or because of damage to DNA caused by the environmental exposures. Cancer causing environmental exposures include substances, such as chemicals in tobacco smoke, radiation such as ultraviolet (UV) rays from the sun, alcohol use and also unhealthy diet.

Besides that, ageing is another fundamental factor for developing cancer. The incidence of cancer rises dramatically with age, mostly due to a build-up risk for specific cancers that increase with age. The overall risk accumulation is combined with the tendency for cellular repair mechanisms to be less effective as a person grows older. There are different types of cancers. However, the most common causes of cancer death are cancers of lung, colorectal, stomach, liver and breast.

In United States, the death rate from cancer has declined steadily over the past 25 years, according to the annual statistic reporting from American Cancer Society. The drop of cancer mortality is mostly due to the reduction of smoking and the advances in early detection and treatment. But not all population group are benefiting. Although the racial gap in the cancer deaths is slowly narrowing, socioeconomic inequalities are widening. Hence, I decided to conduct a study on the cancer cases and deaths to investigate the trend of cancers cases and deaths between the year of 1999 to 2016. Other than that, this study may also alert people on the level of seriousness of cancers to avoid the risk factors of getting cancers.

2.0 FOCUS ON TOPIC

Data collection

The data used for this study is a secondary data from a website. This data is collected using United States Cancer Statistics Data Visualization Tool. This tool makes it easy for anyone to explore and use the latest official federal government cancer data from United States Cancer Statistic (USCS). Data in USCS are used to understand cancer burden and trends recently. Besides, it supports cancer research. Moreover, it measures the progress in cancer control and prevention effects done by the agencies.

From the website, cancer registries collect population-based data about occurrence of cancer, the types of cancer, the site in the body where the cancer first occurred, the extent of disease at time of diagnosis, the planned first course of treatment and the outcome of treatment. For cancer cases, data are reported from variety of medical facilities, including hospital, physicians' officers, radiation facilities, freestanding surgical centers and pathology laboratories. Whereas the death data are recorded on death certificates.

Problem Statement

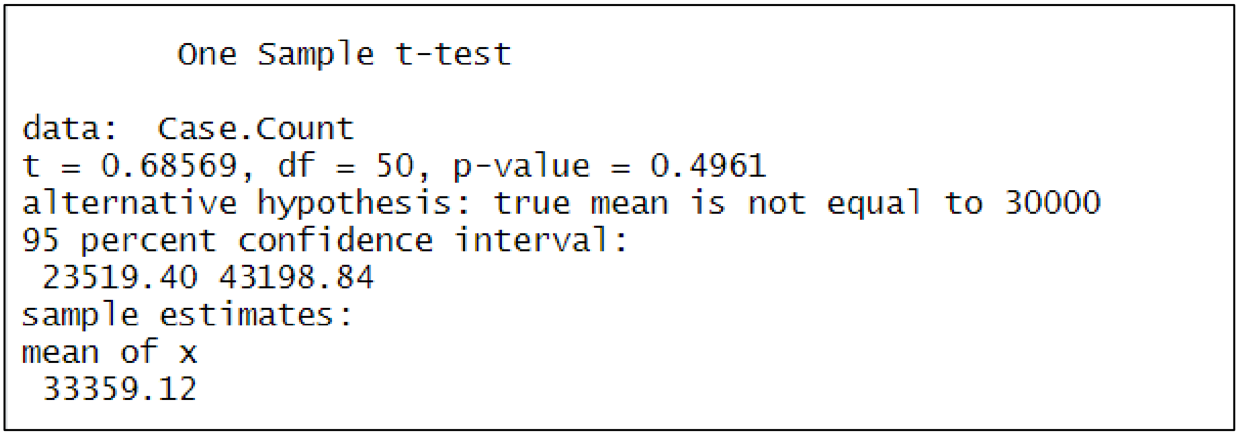
The study of cancers cases and death is to investigate the trend of cancer cases occur for recent year. For the first proposed analysis, the data of the number of new cancers in United States by area is used to estimate the mean (30,000) of new cancer cases in each area. The second statistical analysis was testing the correlation between the number of new cancer cases and cancer deaths. Next, regression test is carried out to identified whether the annual rates of new cancer cases affect the rates of cancer deaths between the year of 1999 and 2016. Apart from that, a chi-square test of independence is used to verify the relationship between number of cancer deaths by gander and races. Finally, the means of the number of cancer cases associated with different risk factor (alcohol, obesity and tobacco) is compared using ANOVA test. All of the test statistic analysis is carried out using R language.

3.0 SUPPORT ON TOPIC

STATISTICAL ANALYSIS

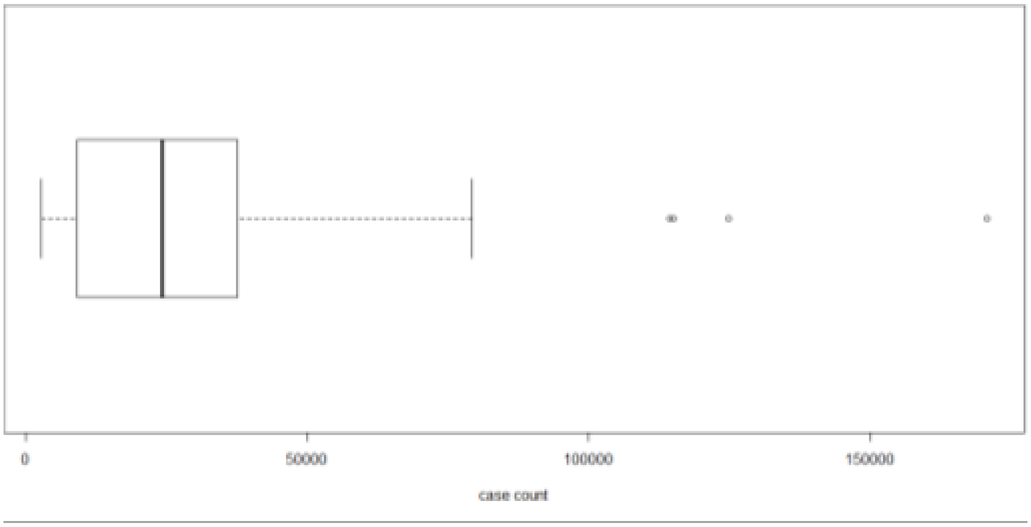
3.1 Hypothesis Testing 1 sample

Number of new cancers in United States by area is used for a 1 sample hypothesis testing. It is used to claim that the mean number of new cancer cases for each area in United States is 30,000 cases at 5% significance level.



```
        One Sample t-test

data:  Case.Count
t = 0.68569, df = 50, p-value = 0.4961
alternative hypothesis: true mean is not equal to 30000
95 percent confidence interval:
 23519.40 43198.84
sample estimates:
mean of x 
 33359.12 
```

- Box plot



- Null hypothesis, H_0 : $\mu = 30000$

  Alternative hypothesis, H_1: $\mu \neq 30000$

- Mean, $\overline{X}$ = 33359.12

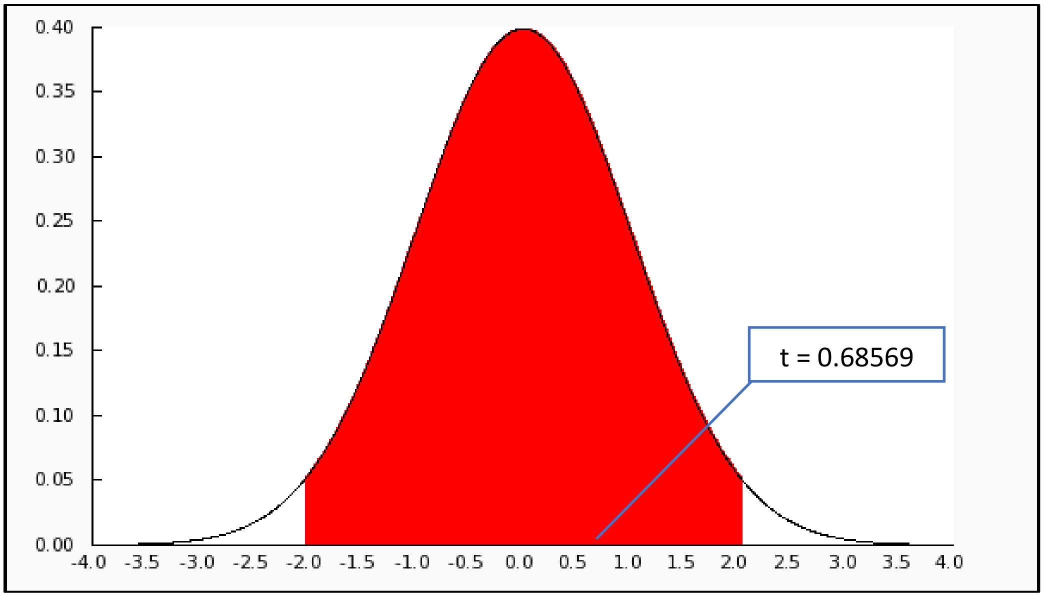
- Level of significance, α = 0.05

  Degree of freedom = 51-1 = 50

- Test statistic, t = 0.68569

  p-value = 0.4961

- t distribution graph

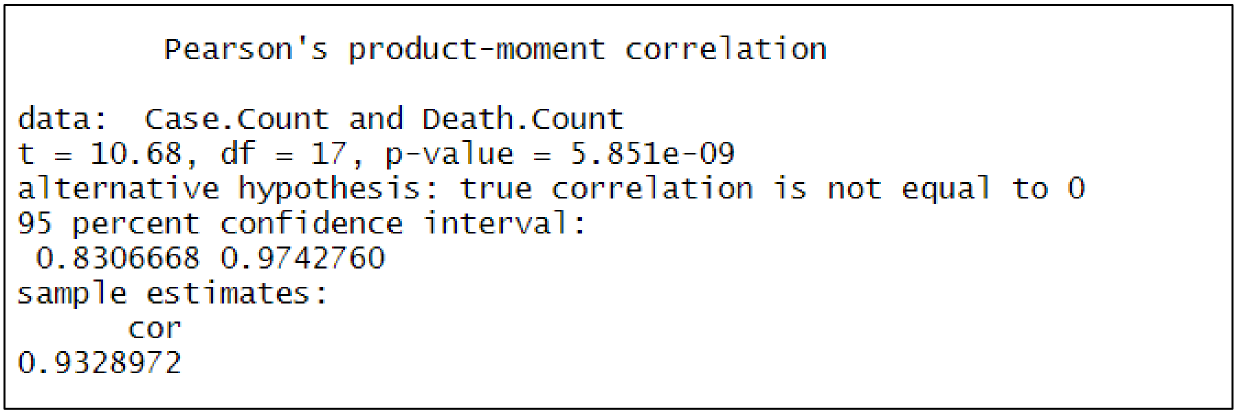


- Conclusion

  For p-value, since 0.4961 > 0.05 and for critical region the value of test statistic, t = 0.68569 do not fall in the rejection region. Thus, do not reject the null hypothesis. There is sufficient evidence to claim that the number of new cancers by area in United States is 30000.

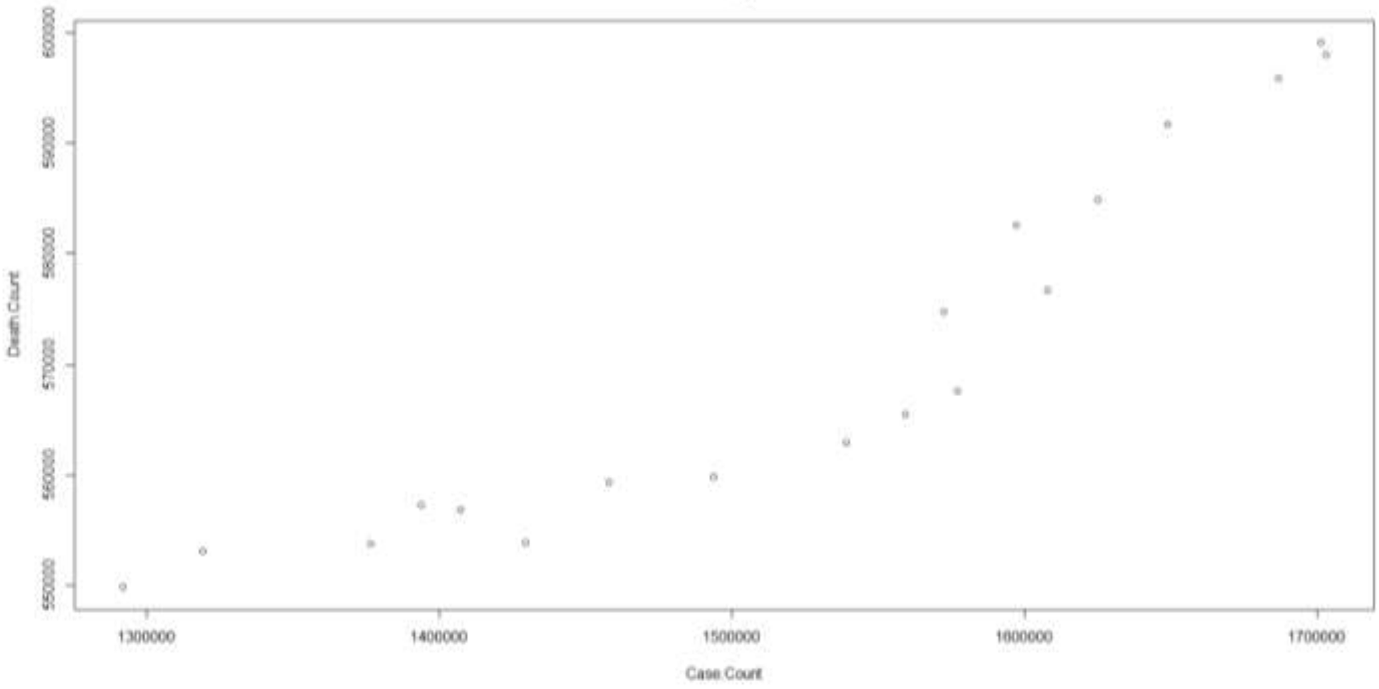
3.2 Correlation

Pearson’s product-moment correlation coefficient is used to conduct the statistical test analysis because both of the data chosen was ratio types. The annual number of cancer deaths and new cancer cases is tested to identify the exits of linear relationship between them at 5% significance level.



```
        Pearson's product-moment correlation

data:  Case.Count and Death.Count
t = 10.68, df = 17, p-value = 5.851e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8306668 0.9742760
sample estimates:
      cor
0.9328972
```

- Scatter plot

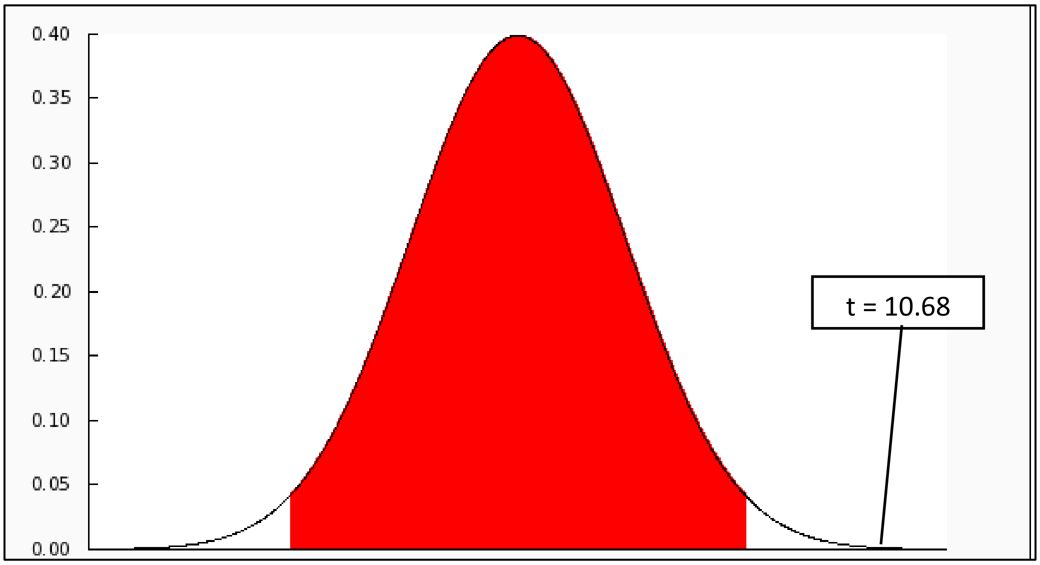


- Null hypothesis, H_0 : p = 0 (no linear correlation)

  Alternative hypothesis, H_1: p ≠ 0 (linear correlation exits)
- Pearson's product-moment correlation coefficient, r = 0.9328972

- Level of significance, α = 0.05

  Degree of freedom = 19-2 = 17
- Test statistic = 10.68
- p-value = 5.851e-09
- T distribution graph

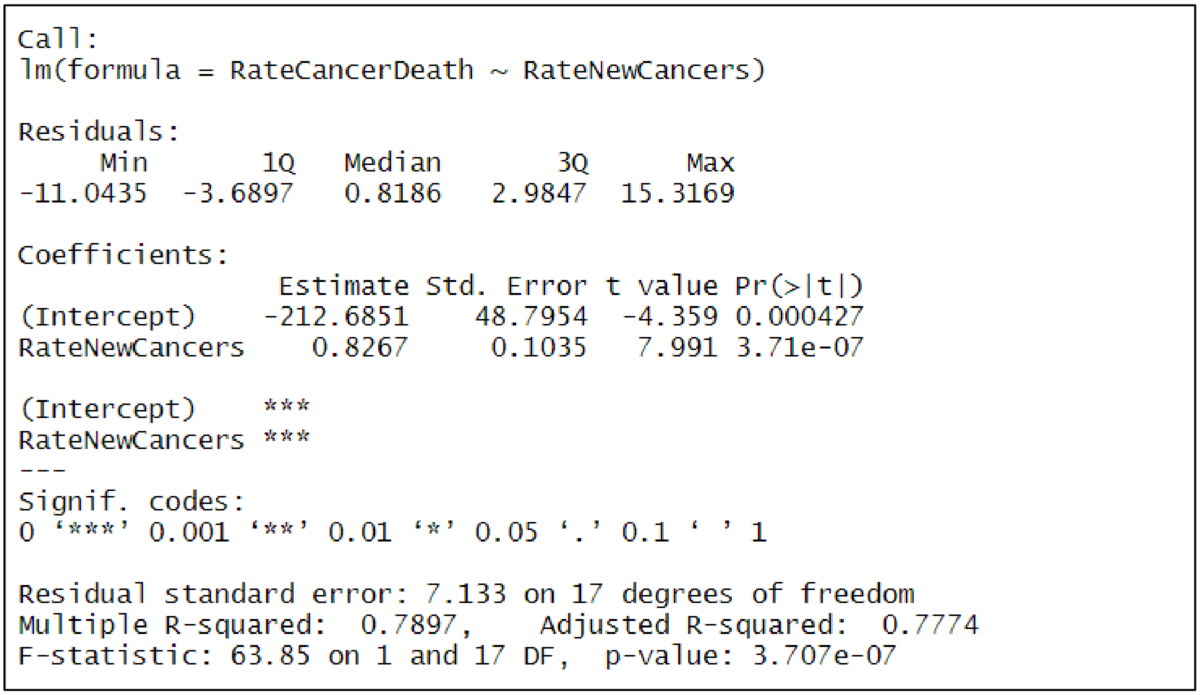


- Conclusion

  For p-value, since 5.851e-09 < 0.05 and for critical region, the value of test statistic, t = 10.68 falls in the rejection region. Thus, the null hypothesis is rejected. There is sufficient evidence to claim that there exists a linear relationship between the number of cancer death and the number of new cancer cases at 5% significance level in United States.

3.3 Regression

The data of the annual rates of cancer deaths and new cancer cases between years of 1999 and 2016 is chosen to propose a regression test. This test is carried out with 5% significance level to identify is there reasonable to conclude that the annual rates of new cancer cases affect the annual rates of cancer deaths in United States.

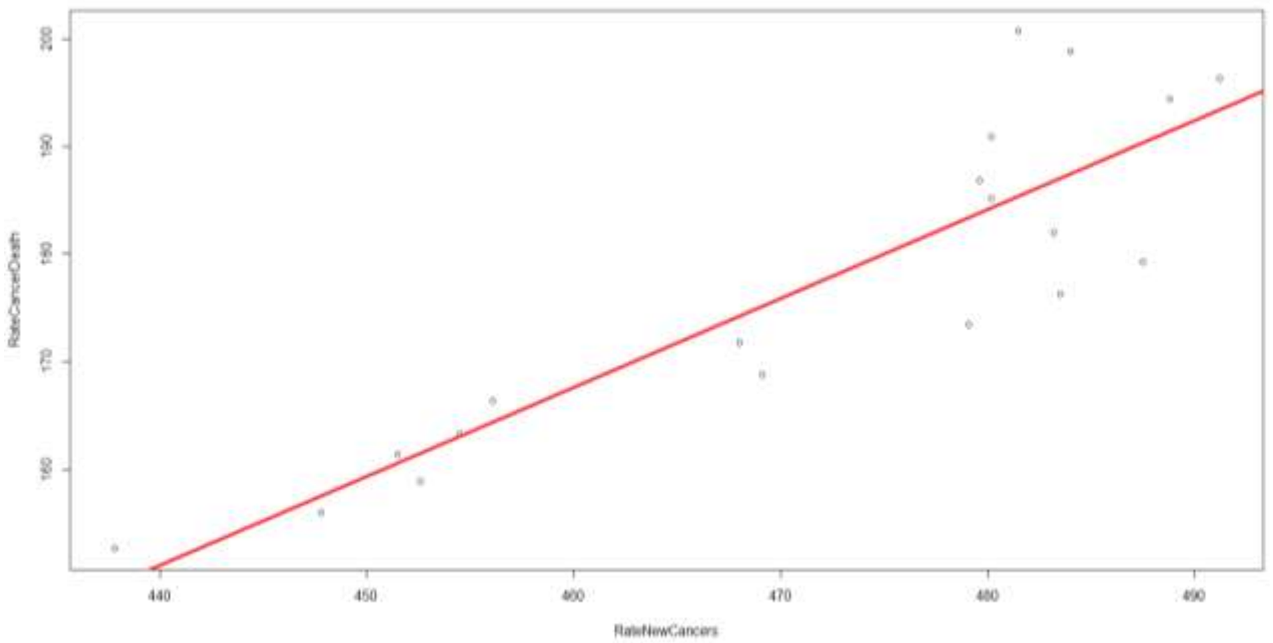


```
Call:
lm(formula = RateCancerDeath ~ RateNewCancers)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.0435  -3.6897   0.8186   2.9847  15.3169 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -212.6851    48.7954  -4.359 0.000427
RateNewCancers    0.8267     0.1035   7.991 3.71e-07
                   
(Intercept)    ***
RateNewCancers ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.133 on 17 degrees of freedom
Multiple R-squared:  0.7897,	Adjusted R-squared:  0.7774 
F-statistic: 63.85 on 1 and 17 DF,  p-value: 3.707e-07
```

- Dependent variable, y : Annual rates of cancer deaths

  Independent variable, x: Annual rates of new cancers

- Regression model



- Correlation coefficient = 0.8886768
- Estimates regression line equation: y = -212.6851+0.8267x
- Intercept, b_0 = -212.6851

  Population slope coefficient, b_1 = 0.8267
- Median = 0.8186
- Null hypothesis, H_0 : the coefficient is equal to 0. (no relationship between x and y)

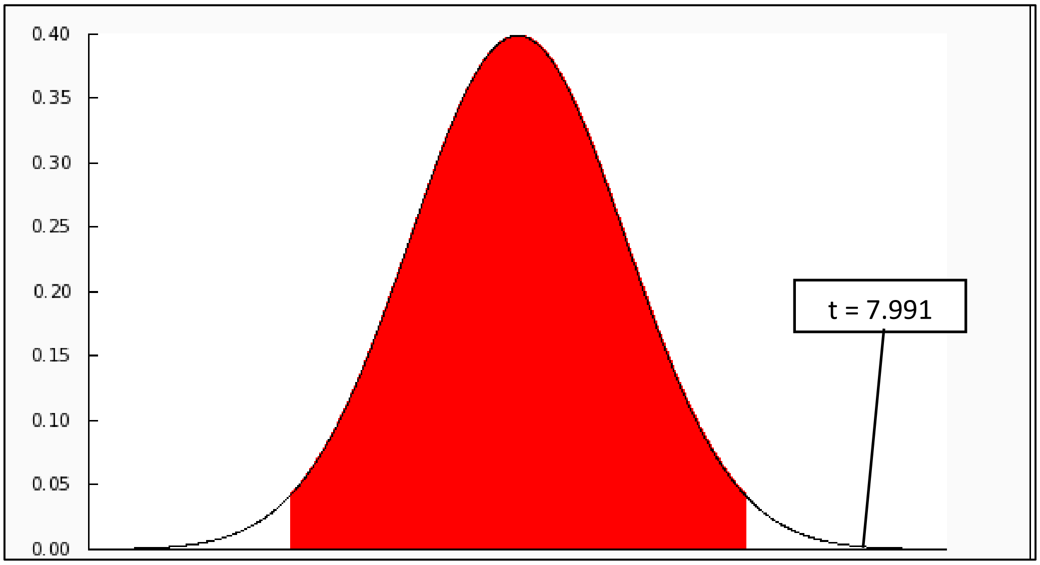
  Alternative hypothesis, H_1: the coefficient is not equal to 0. (there is some relationship between x and y)
- RSE = 7.133

  The observed rates derived from the true regression line by approximately 7.133 units in average.
- R-squared = 0.7897

  It shows 78.97% of the variation in the rate of cancer deaths is explained by variation in the rate of new cancers.
- Level of significance, $\alpha = 0.05$

  Degree of freedom = 19-2 = 17

- Test statistic, t = 7.991
- p-value = 3.707e-07
- t distribution graph

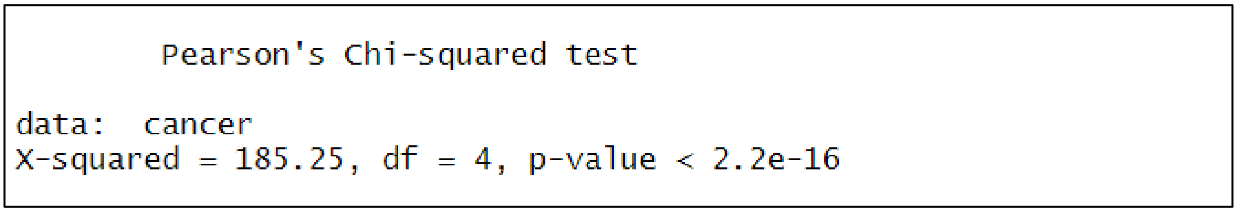


- Conclusion

  For p-value, since 7.414e-07 < 0.05 and for critical region the value of test statistic, t = 7.991 falls in the rejection region. Thus, the null hypothesis is rejected. There is sufficient evidence to claim that annual rates of new cancer cases affect annual rates of cancer deaths.

3.4 Chi square test of independence

Chi square test of independence is carried out to determine if there is a relationship between the number of cancer deaths by races and gender in United States at 5% significance level.



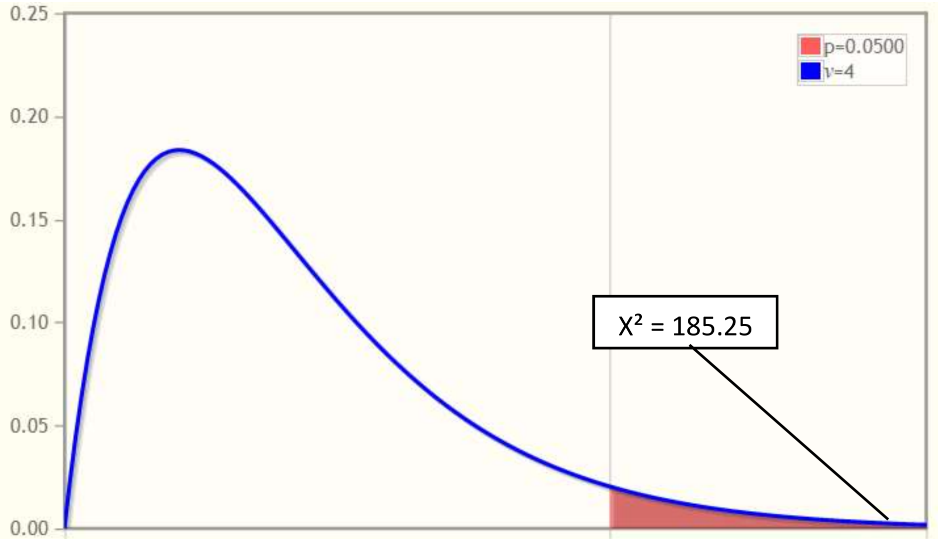
```
        Pearson's Chi-squared test

data:  cancer
X-squared = 185.25, df = 4, p-value < 2.2e-16
```

- Null hypothesis, H_0 : there no relationship between the races and gender of the number of cancer deaths.

  Alternative hypothesis (H1): there is a relationship between the races and gender of the number of cancer deaths.

- Test statistic, $X^2 = 185.25$
- p-value = 2.2e-16
- Level of significance, $\alpha = 0.05$

  Degree of freedom = 4
- chi square graph



- Conclusion

  For p-value, since 2.2e-16 < 0.05 and for critical region the value of test statistic, $X^2 = 185.25$ falls in the rejection region. Thus, the null hypothesis is rejected. There is sufficient evidence to claim that there is a relationship between the number of cancer deaths by gender and races

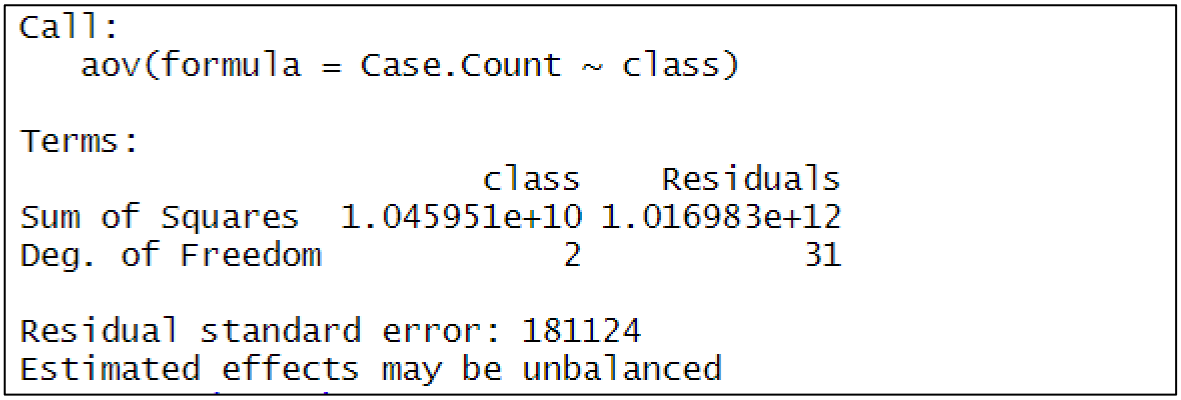
3.5 ANOVA

ANOVA test is used to compare the mean of number of death cancer cases for different risk factors by cancer types. The risk factor is label as A,B and C.

A = Obesity

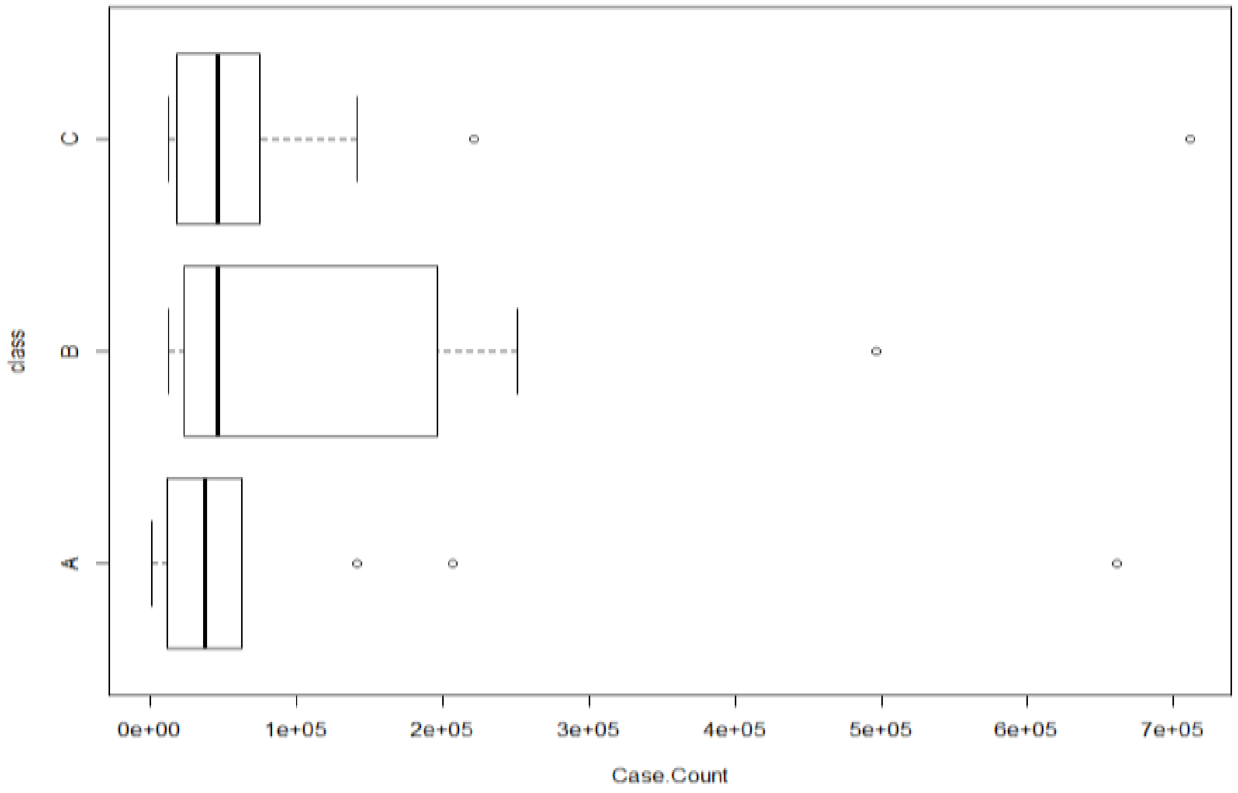
B = Alcohol

C = Tobacco



```
Call:
   aov(formula = Case.Count ~ class)

Terms:
                      class    Residuals
Sum of Squares  1.045951e+10 1.016983e+12
Deg. of Freedom            2           31

Residual standard error: 181124
Estimated effects may be unbalanced
```

- Box plot

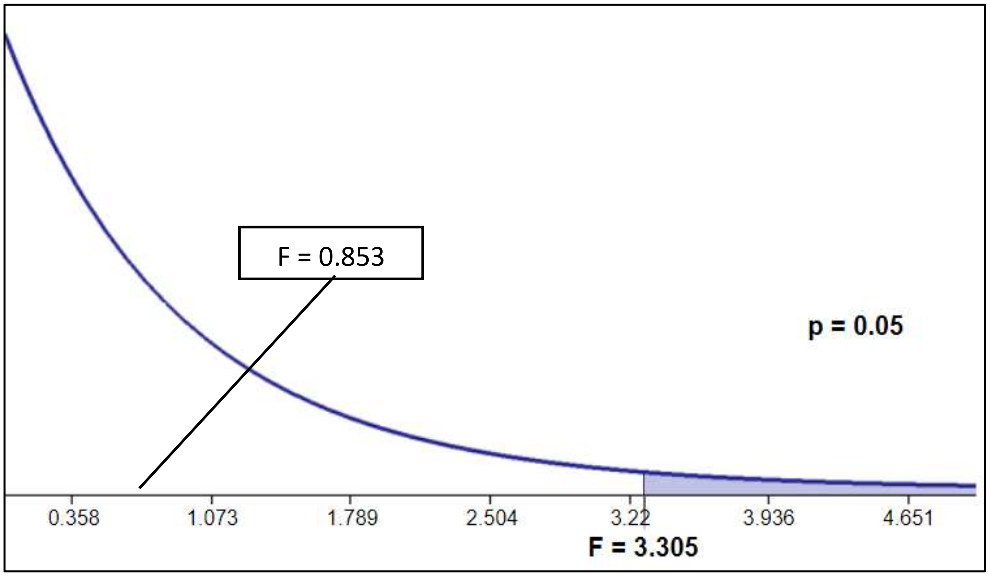


- Null hypothesis, H_0 : $\mu_A = \mu_B = \mu_C$

  Alternative hypothesis, H_1: at least one mean is different
- Test statistic, F = 0.159
- p-value = 0.853
- Level of significance, $\alpha = 0.05$

Degree of freedom: 2, 31

- F distribution graph



- Conclusion

Since p-value = 0.853 > 0.05, thus we do not reject the null hypothesis. There is sufficient evidence to claim that the different risk factors have the same mean for number of cancer case count.

4.0 CONCLUSION

As conclusion, the number of new cancers in a year consider serious. Based on the test statistic done at 5% significance level, I am able to claim that the mean of the new cancers by area in United States was 30,000 cases. Besides, there is a positive linear relationship between the annual number of cancer death and the annual number of new cancers (1999-2016) in United States. Whereas the annual rates of the cancer deaths are affected by the annual rates of new cancers. In some cases, even though the rate is going down, the number of new cases and deaths is going up. This may happen due to the size of population is growing and aging each year. Other than that, I have concluded that the number of cancer deaths by gender (male, female) and the number of cancer deaths by races (White, Black, American Indian/Alaska Native, Asian/Pacific Islander, Hispanic) were dependent. Finally, I have sufficient evidence to claim that the mean for number of cancer deaths by different risk factors (alcohol, tobacco, obesity) were same. Based on the study, the cancer cases are impacted by changes in exposure to risk factors. It shows some of the cancer rates are going down. However, to maintain this situation, strategies and even precaution steps have to take earlier to avoid the risk factors. Besides that, cancer burden can also be reduced through early detection of cancer and management of patients who develop cancer. Many cancers have high chance of cure if diagnosed early and treated adequately.

5.0 REFERENCE

Simon, S. (2019). *Facts & Figures 2019: US Cancer Death Rate has Dropped 27% in 25 Years.* American Cancer Society.

USCS. (2020, June). *United States Cancer Statistics: Data Vizualizations*. Retrieved from Centers of disease control and prevention (CDC): https://gis.cdc.gov/Cancer/USCS/DataViz.html

world health organization. (2018, september 12). Retrieved from cancer: https://www.who.int/news-room/fact-sheets/detail/cancer