**Use R to Analyze the Data related to Risk Factors associated with Cancer cases in the United States in 2017**

**Statistical Computing-Project 1**

**Objective:** The purpose of this project is to examine and study alcohol-associated cancer cases in the United States during 2017. I will analyze multiple data sets for this risk. The datasets that will be examined for alcohol-associated risk are cancer cases the age groups (Age), type of cancer (CancerType), cancer cases by state(Map), and cancer cases by race. I want to know what age group is more at risk for alcohol-associated cancer, the types of cancers that are prevalent with alcohol-association, and states that suffer the most from these type of cancer cases. The variables that will be analyzed are of alcohol-associated cancer per 100,000 people (Rate), gender of people which will be both male and female for all the data (Sex), case count of risk associated cancer (Count), population that is being looked at (Population) . The data sets also include the lower and upper confidence intervals at a 95% confidence level(lci & uci).

I will start by analyzing the rate of risk-associated cancer cases related to age for alcohol-associated risks.

**Alcohol-associated**

> alcoholAge = read.csv("C:\\Users\\Cellus\\OneDrive\\Documents\\Fall 2020\\Statistical Computing - STAT 40001\\Project 1\\Alcohol-associated Cancers Data\\USCS\_RFAge-Alcohol.csv",header=T)

> ls(alcoholAge)

[1] "Age" "CancerType" "Count" "lci" "Population"

[6] "Race" "Rate" "Sex" "uci"

I used barplot comparing the rate of cancer cases related to alcohol by age groups for all genders and races.

> attach(alcoholAge)

> barplot(Rate~Age, main="Alcohol-associated cancer rate per 100,000 people by age",xlab="Age (in Years)", ylab="Rate (per 100,00 people)")



This model shows alcohol-associated cancer rates by age group:

> m = lm(Rate~-1+Age)

> m

Call:

lm(formula = Rate ~ -1 + Age)

Coefficients:

Age'<20' Age'20-24' Age'25-29' Age'30-34' Age'35-39' Age'40-44'

1.0 3.7 9.8 23.3 46.9 91.8

Age'45-49' Age'50-54' Age'55-59' Age'60-64' Age'65-69' Age'70-74'

151.4 217.3 272.2 350.5 446.0 493.9

Age'75-79' Age'80-84' Age'85+'

535.5 555.9 518.7

This boxplot compares the number of alcohol-associated cancer cases by age.

> barplot(Count~Age, main="# of Alcohol-associated cases by age group")



> AgeCount

CancerType Race Sex Age Count Population

1 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '<20' 847 82066030

2 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '20-24' 815 22065664

3 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '25-29' 2281 23336175

4 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '30-34' 5104 21938443

5 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '35-39' 9935 21190454

6 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '40-44' 17980 19590752

7 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '45-49' 31673 20916186

8 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '50-54' 46390 21347575

9 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '55-59' 59779 21959910

10 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '60-64' 69897 19942009

11 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '65-69' 74944 16803000

12 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '70-74' 63363 12829683

13 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '75-79' 46752 8730528

14 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '80-84' 33131 5960298

15 'All Alcohol-associated Cancers' 'All Races' 'Male and Female' '85+' 33559 6470414

This model shows alcohol-associated cancer cases by age group; intercept added by predictor for each group:

> m = lm(Count~Age)

> m

Call:

lm(formula = Count ~ Age)

Coefficients:

(Intercept) Age'20-24' Age'25-29' Age'30-34' Age'35-39'

847 -32 1434 4257 9088

Age'40-44' Age'45-49' Age'50-54' Age'55-59' Age'60-64'

17133 30826 45543 58932 69050

Age'65-69' Age'70-74' Age'75-79' Age'80-84' Age'85+'

74097 62516 45905 32284 32712

So, reject the null hypothesis, meaning that the proportions are not equal.

I will use two proportion sample test to examine the closeness of age group “65-69” and “85+”.

> prop.test(x=c(74944,33559),n=c(16803000,6470414),alternative="two.sided")

2-sample test for equality of proportions with continuity

correction

data: c(74944, 33559) out of c(16803000, 6470414)

X-squared = 531.01, df = 1, p-value < 2.2e-16

alternative hypothesis: two.sided

95 percent confidence interval:

-0.0007903438 -0.0006624055

sample estimates:

prop 1 prop 2

0.004460156 0.005186531

So, reject the null hypothesis, meaning that the proportions are not equal.

I will do one more test using ages “75-79” and ages “80-84” due to their rates being the closest between any set of age groups.

> prop.test(x=c(46752,33131),n=c(8730528,5960298),alternative="two.sided")

2-sample test for equality of proportions with continuity

correction

data: c(46752, 33131) out of c(8730528, 5960298)

X-squared = 27.116, df = 1, p-value = 1.916e-07

alternative hypothesis: two.sided

95 percent confidence interval:

-0.0002806054 -0.0001266188

sample estimates:

prop 1 prop 2

0.005355003 0.005558615

So, reject the null hypothesis, meaning that the proportions are again not equal. I will now test this to see if the proportion for ages “75-79” is less than the proportion for ages “80-84”.

> prop.test(x=c(46752,33131),n=c(8730528,5960298),alternative="less")

2-sample test for equality of proportions with continuity

correction

data: c(46752, 33131) out of c(8730528, 5960298)

X-squared = 27.116, df = 1, p-value = 9.58e-08

alternative hypothesis: less

95 percent confidence interval:

-1.0000000000 -0.0001389746

sample estimates:

prop 1 prop 2

0.005355003 0.005558615

Reject null hypothesis. So, it is in fact true that the proportion of alcohol-associated cases for the ages “75-79” is less than that of the ages “80-84”. This is proven with the rate of “80-84” being higher than the rate of “75-79”.

Next, I will examine the rate of new alcohol-associated cancers by cancer type.

> alcoholType = read.csv("C:\\Users\\Cellus\\OneDrive\\Documents\\Fall 2020\\Statistical Computing - STAT 40001\\Project 1\\Alcohol-associated Cancers Data\\USCS\_RFCancerSite-Alcohol.csv",header=T)

> attach(alcoholType)

This boxplot shows the rate of new alcohol-associated cancer cases. This is out of the population of America in 2017 which 325 million. For female breast cancer, it is out of only the female population which is 165 million.

> CancerType = alcoholType[,c("CancerType","Race","Sex","Rate")]

> CancerType

CancerType Race Sex Rate

1 All Cancers' 'All Races' 'Male and Female' 128.9

2 'Female Breast' 'All Races' 'Male and Female' 125.1

3 'Colon and Rectum' 'All Races' 'Male and Female' 36.8

4 Lip, Oral Cavity and Pharynx' All Races' 'Male and Female' 11.7

5 'Liver' 'All Races' 'Male and Female' 6.8

6 'Esophagus' 'All Races' 'Male and Female' 4.4

7 'Larynx' 'All Races' 'Male and Female' 3.0

> barplot(Rate~CancerType, main="Rate of new alcohol-associated cancers by cancer type", xlab= "Cancer Type", ylab= "Rate per 100,000 people", col= c(2,3,4,5,6,7,8))



I will analyze total number of cancer cases for the population by cancer type.

> barplot(Count~alcoholType$CancerType, main="# of alcohol-associated cancers by cancer type", xlab= "Cancer Type", ylab= "Case Count", col= c(2,3,4,5,6,7,8))



The next part of the data that I will analyze is the relationship between alcohol-associated cancer cases and races of people. Analyzing cases and the rate of these new cancer cases.

First, I will analyze the rate of these new cancer cases by race using a boxplot and pie chart.

> alcoholRace = read.csv("C:\\Users\\Cellus\\OneDrive\\Documents\\Fall 2020\\Statistical Computing - STAT 40001\\Project 1\\Alcohol-associated Cancers Data\\USCS\_RFRace-Alcohol.csv",header=T)

> alcoholRace

CancerType Race Sex Rate

1 'All Alcohol-associated Cancers' 'Asian/Pacific Islander' 'Male and Female' 102.1

2 'All Alcohol-associated Cancers' 'American Indian/Alaska Native' 'Male and Female' 86.3

3 'All Alcohol-associated Cancers' 'Non-Hispanic' 'Male and Female' 132.4

4 'All Alcohol-associated Cancers' 'Hispanic' 'Male and Female' 105.2

5 'All Alcohol-associated Cancers' 'Black' 'Male and Female' 132.8

6 'All Alcohol-associated Cancers' 'White' 'Male and Female' 128.0

lci uci Count Population

1 '100.8' '103.5' 21420 21130607

2 '83.3' '89.4' 3361 4684666

3 '132.0' '132.8' 451851 266446132

4 '104.2' '106.3' 44587 58700989

5 '131.7' '134.0' 57606 45788656

6 '127.6' '128.4' 406508 253543192

> attach(alcoholRace)

> pie(Rate,Race, main ="Rate of alcohol-associated cancers by race")



> barplot(Rate~Race, main="Rate of alcohol-assoicated cancers by race", ylim=c(60,140), beside=T, xpd=F, col=c(2,3,4,5,6,7))



I will analyze the number of cases compared to the population for each race.

> casePop = cbind(Race,Count,Population)

> casePop

Race Count Population

[1,] "'Asian/Pacific Islander'" "21420" "21130607"

[2,] "'American Indian/Alaska Native'" "3361" "4684666"

[3,] "'Non-Hispanic'" "451851" "266446132"

[4,] "'Hispanic'" "44587" "58700989"

[5,] "'Black'" "57606" "45788656"

[6,] "'White'" "406508" "253543192"

The next piece of data we will analyze is the alcohol associated cancers by state. We will use bar plots to show the rates per 100,000 people for each state and the population for each state.

> rate = scan()

1: 133.5

2: 124.5

3: 109.4

4: 132

5: 119.8

6: 123.8

7: 132.7

8: 126.7

9: 120.8

10: 133.6

11:

Read 10 items

> names(rate)=c("AL","AK","AZ","AR","CA","CO","CT","DE","FL","GA")

> barplot(rate,col=c(1,2,3,4,5,6,7,8,9,10), main = "Rate of new cacners cases in Alabama-Georgia")





> rate = scan()

1: 140.7

2: 127.3

3: 136.7

4: 130.3

5: 140.9

6: 133.1

7: 142.2

8: 143.5

9: 131

10: 130.9

11:

Read 10 items

> names(rate)=c("HI","ID","IL","IN","IA","KS","KY","LA","ME","MD")

> barplot(rate,col=c(1,2,3,4,5,6,7,8,9,10), ylim=c(80,150),beside=T, xpd=F, main = "Rate of new cancers cases in Hawaii-Maryland")

> pop=scan()

1: 1424203

2: 1718904

3: 12786196

4: 6660082

5: 3143637

6: 2910689

7: 4453874

8: 4670818

9: 1335063

10: 6024891

11:

Read 10 items

> names(pop)=c("HI","ID","IL","IN","IA","KS","KY","LA","ME","MD")

> barplot(pop,col=c(1,2,3,4,5,6,7,8,9,10), main = "Population of Hawaii-Maryland")



> rate = scan()

1: 130.4

2: 125.7

3: 134.9

4: 141.6

5: 137.7

6: 134.6

7: 134.2

8: 118.6

9: 128.2

10: 138

11:

Read 10 items

> names(rate)=c("MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ")

> barplot(rate,col=c(1,2,3,4,5,6,7,8,9,10), ylim=c(80,150), beside=T, xpd=F, main = "Rate of new cancers cases in Massachusetts-New Jersey")



> pop = scan()

1: 6863246

2: 9976447

3: 5568155

4: 2989663

5: 6108612

6: 1053090

7: 1917575

8: 2972405

9: 1349767

10: 8888543

11:

Read 10 items

> names(pop)=c("MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ")

> barplot(pop,col=c(1,2,3,4,5,6,7,8,9,10), main = "Population of Massachusetts-New Jersey")



> rate = scan()

1: 113.9

2: 135.5

3: 134

4: 133.1

5: 135.9

6: 133

7: 121.1

8: 136.3

9: 129

10: 132

11:

Read 10 items

> names(rate)=c("NM","NY","NC","ND","OH","OK","OR","PA","RI","SC")

> barplot(rate,col=c(1,2,3,4,5,6,7,8,9,10), ylim=c(80,150), beside=T, xpd=F, main = "Rate of new cancers cases in New Mexico-South Carolina")

> pop = scan()

1: 2093395

2: 19590719

3: 10270800

4: 755176

5: 11664129

6: 3932640

7: 4146592

8: 12790447

9: 1056486

10: 5021219

11:

Read 10 items

> names(pop)=c("NM","NY","NC","ND","OH","OK","OR","PA","RI","SC")

> barplot(pop,col=c(1,2,3,4,5,6,7,8,9,10), main = "Population of New Mexico-South Carolina")



> rate = scan()

1: 130.6

2: 130.6

3: 124

4: 105.9

5: 124

6: 122.1

7: 130.2

8: 136.7

9: 128.5

10: 111.4

11:

Read 10 items

> names(rate)=c("SD","TN","TX","UT","VT","VA","WA","WV","WI","WY")

> barplot(rate,col=c(1,2,3,4,5,6,7,8,9,10), ylim=c(80,150), beside=T, xpd=F, main = "Rate of new cancers cases in South Dakota-Wyoming")



> pop = scan()

1: 873286

2: 6708794

3: 28322717

4: 3103118

5: 624525

6: 8465207

7: 7425432

8: 1817048

9: 5792051

10: 578934

11:

Read 10 items

> names(pop)=c("SD","TN","TX","UT","VT","VA","WA","WV","WI","WY")

> barplot(pop,col=c(1,2,3,4,5,6,7,8,9,10), main = "Population of South Dakota-Wyoming")



I tested the proportions between Illinois and West Virginia as they both have the same rate at 136.7 per 100,000 people. Also note that the rate is an age-adjusted rate.

First, I tested for equality.

> prop.test(x=c(20514,3396), n=c(12786196,1817048))

2-sample test for equality of proportions with continuity

correction

data: c(20514, 3396) out of c(12786196, 1817048)

X-squared = 67.97, df = 1, p-value < 2.2e-16

alternative hypothesis: two.sided

95 percent confidence interval:

-0.0003314146 -0.0001977435

sample estimates:

prop 1 prop 2

0.001604386 0.001868965

Reject null hypothesis so the proportions are not equal. Prop 1 (Illinois) is shown to have a smaller proportion. So I test to see if Illinois’s proportion is less than West Virginia’s proportion.

> prop.test(x=c(20514,3396), n=c(12786196,1817048), alternative="less")

2-sample test for equality of proportions with continuity

correction

data: c(20514, 3396) out of c(12786196, 1817048)

X-squared = 67.97, df = 1, p-value < 2.2e-16

alternative hypothesis: less

95 percent confidence interval:

-1.0000000000 -0.0002084383

sample estimates:

prop 1 prop 2

0.001604386 0.001868965

Reject null hypothesis. So, it is in fact true that Illinois’s has a smaller proportion of alcohol associated cancer cases compared to West Virginia even though their “age-adjusted” rates are the same.

I will now test whether the state’s rates are higher than 100 people.

> map = read.csv("C:\\Users\\Cellus\\OneDrive\\Documents\\Fall 2020\\Statistical Computing - STAT 40001\\Project 1\\Alcohol-associated Cancers Data\\USCS\_RFMap-Alcohol.csv",header=T)

> t.test(map$Rate, alt="greater", mu=100)

One Sample t-test

data: map$Rate

t = 25.642, df = 50, p-value < 2.2e-16

alternative hypothesis: true mean is greater than 100

95 percent confidence interval:

127.7461 Inf

sample estimates:

mean of x

129.6863

There is enough evidence to reject the null hypothesis and support that the alcohol associated cancer rates are more than 100 people per 100,000. Next, I generated the 95% confidence interval for the rates.

> t.test(map$Rate)

One Sample t-test

data: map$Rate

t = 109.82, df = 49, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

127.3287 132.0753

sample estimates:

mean of x

129.702

Confidence interval falls between 127.3287 and 132.0753.

Analyzing the cancer rate of each state in the dataset, we find that the state with the largest rate of alcohol-associated cancers is Louisiana at 143.5.

> map = read.csv("C:\\Users\\Cellus\\OneDrive\\Documents\\Fall 2020\\Statistical Computing - STAT 40001\\Project 1\\Alcohol-associated Cancers Data\\USCS\_RFMap-Alcohol.csv",header=T)

> newMap=map[rev(order(map[,"Rate"])),]

> head(newMap)

CancerType Area Race Sex Rate lci uci Count Population

18 'All Alcohol-associated Cancers' Louisiana 'All Races' 'Male and Female' 143.5 '140.2' '146.8' 7755 '4670818'

17 'All Alcohol-associated Cancers' Kentucky 'All Races' 'Male and Female' 142.2 '138.9' '145.5' 7705 '4453874'

24 'All Alcohol-associated Cancers' Mississippi 'All Races' 'Male and Female' 141.6 '137.6' '145.7' 4997 '2989663'

15 'All Alcohol-associated Cancers' Iowa 'All Races' 'Male and Female' 140.9 '137.0' '144.8' 5404 '3143637'

11 'All Alcohol-associated Cancers' Hawaii 'All Races' 'Male and Female' 140.7 '135.0' '146.6' 2480 '1424203'

30 'All Alcohol-associated Cancers' New Jersey 'All Races' 'Male and Female' 138.0 '135.7' '140.3' 14961 '8888543'

**Conclusion:** I have learned much about my analysis of alcohol-associated cancer cases in America. I have learned that the primary age group where the risk of alcohol-associated cancer is high is within the 60+ years age group. Next dataset analyzed was the type of cancers that are alcohol-associated. I have learned that female breast cancer has the highest rate by cancer type. The next largest is colon/rectum cancer. The next variable was cancer cases by race which I found to be very interesting and may want to look at more closely also including gender. The groups with the highest rate of alcohol-associated cancers were white and non-Hispanic. Lastly, I analyzed the cancer cases by state in America. The state with the highest cancer rate is Louisiana.

This I would also like to analyze for future work within states. I would like to go back to my source of this data which was CDC (Center for Disease Control), to study and analyze these states with high cancer rate with respect to gender and race. This can be future work for my final project.