The analysis based on following questions:

1. Does the elderly person die more frequently due to COVID? Justify your answer based on t-test results.

2. Does the male get more affected due to COVID as compared to female? Justify your answer based on t-test results.

3. If you consider two groups of people who were from Wuhan and were visiting Wuhan which group of people were more affected due to COVID?

4. Consider two countries China and South Korea. Answer the above three questions based on t-test results.

Answer to the question no 1:

Code:

```
# Load the dataset

rm(list = ls())

library(Hmisc)


data <- read.csv("E:/4_10_2021/COVID-19/COVID19_line_list_data.csv")

describe(data)


data$death_dummy <- as.integer(data$death != 0)

unique(data$death_dummy)


sum(data$death_dummy) / nrow(data)


dead = subset(data, death_dummy == 1)

alive = subset(data, death_dummy == 0)


mean(dead$age, na.rm = TRUE)

mean(alive$age, na.rm = TRUE)

t.test(alive$age, dead$age, alternative = "two.sided", conf.level = 0.99)
```

Result:

```
> mean(dead$age, na.rm = TRUE)

[1] 68.58621

> mean(alive$age, na.rm = TRUE)

[1] 48.07229
```

```
p-value < 2.2e-16

99 percent confidence interval:

 -25.52122 -15.50661
```

Analysis: After doing the t-test we can see that in 99% confidence interval there is 99% chance that the difference between the people who alive and dead in age is from 25 years to 16 years. So, on average the person who alive is much younger. Also if we look at the p-value which is 2.2e-16 means almost 0. So there is 0 percent chance that we have got such an extreme result at random from this sample. So, indeed we can reject the hypothesis because the standard significant threshold is 0.05. So in conclusion we can say that after doing the t-test that the elderly person died more frequently in COVID-19.

Answer to the question no 2:

Code:

```
men = subset(data, gender == "male")

women = subset(data, gender == "female")
```

```
mean(men$death_dummy, na.rm = TRUE)

mean(women$death_dummy, na.rm = TRUE)


t.test(men$death_dummy,women$death_dummy, alternative = "two.sided", conf.level = 0.99)
```

Resut:

```
> mean(men$death_dummy, na.rm = TRUE)

[1] 0.08461538

> mean(women$death_dummy, na.rm = TRUE)

[1] 0.03664921


99 percent confidence interval:

 0.007817675 0.088114665

p-value = 0.002105
```

Analysis

As we have seen the mean of the male death rate is 8.5% compared to female death rate of 3.5%. So, from the t-test with 99% confidence interval we found that male has 0.7% to 8.8% higher change of death than female. Also our p-value comes around 0.002, which much less than the standard significant threshold of 0.05. So, as a result we can claim that male get more affected due to COVID compared to female.

Answer to the question no 3:

Code:

```
# from Wuhan

fw = subset(data, from.Wuhan == 1)

fws = sum(fw$death_dummy, na.rm = TRUE)

fws


# Visiting Wuhan

vw = subset(data, visiting.Wuhan == 1)

vws = sum(vw$death_dummy, na.rm = TRUE)

vws
```

Result:

```
> fws

[1] 34

> vws

[1] 1
```

Analysis: From the dataset we have found that 34 peoples got affected who were from Wuhan compared to 1 person who was visiting Wuhan. So we clearly say that people from Wuhan got more affected than the people were visiting Wuhan.

Answer to the question no 4:

Code:

```
china = subset(data, country == "China")

south_korea = subset(data, country == "South Korea")
```

For China:

Answer to the question no 1C:

Code:

```
china$death_dummy <- as.integer(china$death != 0)

unique(china$death_dummy)


sum(china$death_dummy) / nrow(china)


dead = subset(china, death_dummy == 1)

alive = subset(china, death_dummy == 0)


mean(dead$age, na.rm = TRUE)
```

```
mean(alive$age, na.rm = TRUE)


t.test(alive$age, dead$age, alternative = "two.sided", conf.level = 0.99)
```

Result:

```
> mean(dead$age, na.rm = TRUE)

[1] 71.05128

> mean(alive$age, na.rm = TRUE)

[1] 43.30464


99 percent confidence interval:

 -33.73155 -21.76174

p-value < 2.2e-16
```

Analysis: After doing the t-test we can see that in 99% confidence interval there is 99% chance that the difference between the people who alive and dead in age is from 33 years to 21 years. So, on average the person who alive is much younger. Also if we look at the p-value which is 2.2e-16 means almost 0. So there is 0 percent chance that we have got such an extreme result at random from this sample. So, indeed we can reject the hypothesis.

Answer to the question no 2C:

Code:

```
men = subset(china, gender == "male")

women = subset(china, gender == "female")
```

```
mean(men$death_dummy, na.rm = TRUE)

mean(women$death_dummy, na.rm = TRUE)


t.test(men$death_dummy,women$death_dummy, alternative = "two.sided", conf.level = 0.99)
```

Result:

```
> mean(men$death_dummy, na.rm = TRUE)

[1] 0.2177419

> mean(women$death_dummy, na.rm = TRUE)

[1] 0.1643836

99 percent confidence interval:

 -0.09618598  0.20290273


p-value = 0.03538
```

Analysis: As we have seen the mean of the male death rate is 21.7 % compared to female death rate of 16.4%. So, from the t-test with 99% confidence interval we found that male has 9.6% to 20.2% higher change of death than female. Also our p-value comes around 0.035, which much less than the standard significant threshold of 0.05. So, as a result we can claim that male get more affected due to COVID compared to female.

Answer to the question the no 3C:

Code:

```
# from Wuhan

fw = subset(china,from.Wuhan == 1)

fws = sum(fw$death_dummy, na.rm = TRUE)

fws


# Visiting Wuhan

vw = subset(china, visiting.Wuhan == 1)

vws = sum(vw$death_dummy, na.rm = TRUE)

vws
```

Result:

```
> fws

[1] 32


> vws

[1] 0
```

Analysis: From the dataset we have found that 32 peoples got affected who were from Wuhan compared to 0 person who was visiting Wuhan. So we clearly say that people from Wuhan got more affected than the people were visiting Wuhan.

For South Korea

Answer to the question no 1S

Code:

```
south_korea$death_dummy <- as.integer(south_korea$death != 0)
unique(south_korea$death_dummy)

sum(south_korea$death_dummy) / nrow(south_korea)

dead = subset(south_korea, death_dummy == 1)
alive = subset(south_korea, death_dummy == 0)

mean(dead$age, na.rm = TRUE)
mean(alive$age, na.rm = TRUE)

t.test(alive$age, dead$age, alternative = "two.sided", conf.level = 0.99)
```

Result:

```
> mean(dead$age, na.rm = TRUE)
[1] 57.66667
```

> mean(alive$age, na.rm = TRUE)

[1] 46.55422

99 percent confidence interval:

-25.614907 -3.390008

p-value = 0.03657

Analysis:

After doing the t-test we can see that in 99% confidence interval there is 99% chance that the difference between the people who alive and dead in age is from 25 years to 3 years. So, on average the person who alive is much younger. Also if we look at the p-value which is 0.03657 means almost 0. So there is 0 percent chance that we have got such an extreme result at random from this sample. So, indeed we can reject the hypothesis.

Answer to the question the no 2S:

Code:

men = subset(south_korea, gender == "male")

women = subset(south_korea, gender == "female")

```
mean(men$death_dummy, na.rm = TRUE)

mean(women$death_dummy, na.rm = TRUE)


t.test(men$death_dummy,women$death_dummy, alternative = "two.sided", conf.level = 0.99)
```

Result:

```
> mean(men$death_dummy, na.rm = TRUE)

[1] 0.1860465

> mean(women$death_dummy, na.rm = TRUE)

[1] 0.02040816


99 percent confidence interval:

 0.003971023  0.335247720


p-value = 0.01176
```

Analysis:

As we have seen man from the mean the male has death rate of 18.6 % compared to female death rate of 2%. So, from the t-test with 99% confidence interval we found that male has 3% to 33% higher change of death than female. Also our p-value comes around 0.035, which much less than the standard significant threshold of 0.01. So, as a result we can claim that male get more affected due to COVID compared to female.

Answer to the question the no 3S:

Code:

```
# from Wuhan

fw = subset(south_korea,from.Wuhan == 1)

fws = sum(fw$death_dummy, na.rm = TRUE)

fws


# Visiting Wuhan

vw = subset(south_korea, visiting.Wuhan == 1)

vws = sum(vw$death_dummy, na.rm = TRUE)

vws
```

Result:

```
 > fws

[1] 0

> vws

[1] 0
```

Result: We have found that no one affected from Wuhan or Visiting Wuhan in case of South Korea data.