IB Mathematics SL

Average Income and COVID-19 Deaths

Is there a relationship between average income and COVID-19 deaths in the United States in 2020?

Word Count: 1607

000877-0116

Table of Contents

I. Rationale……..……3

II. Introduction……..……3

III. Linear Regression…….……..……..……...…5

IV. Pearson correlation coefficient……..…..…….10

V. Conditional Probability…….…..…12

VI. Conclusion…….13

VII. References…….15

Rationale:

The SARS-CoV-2 (COVID-19) pandemic has influenced almost every single life in the world. Individuals have adapted to the situation by quarantining, wearing face masks, and social distancing. Personally, I had to switch to online school over spring and I had to stay at home throughout summer. Since I have a compromised immune system, I would like to know how the COVID-19 can affect me. Additionally, since this virus is very new, the average number of deaths is increasing every day in the United States.

The average income is a vital component of the United States' demographics. Since the COVID-19 has impacted many lives in the United States, there is a chance that each state's average income has been influenced.

It is important to investigate a relationship between the average income and the number of COVID-19 fatalities to understand the spread of the virus throughout the United States. This can aid future researchers who are studying how the average income of each state can impact the death rate in other diseases.

Introduction:

The average income in each state and COVID-19 deaths in the United States will be investigated. This topic will be studied using the mathematical concepts of linear regression and correlation coefficient to determine whether these variables have any relation to each other.

A linear regression line is a model of the relationship between two variables that uses a linear equation to explain the observed data (Yale). To calculate the linear regression line, it is important to look at the formula:

$$y = a + bx$$

a is the y-intercept (the point where the line crosses the x-axis)

b is the slope

The correlation coefficient determines the extent to which two variables influence each other. The Pearson correlation coefficient ranges from -1 to 1. A strong relationship is one that is closer to -1 or 1, and a weak relationship is one that is closer to 0. Moreover, a positive correlation indicates that as one variable increases, the other variable increases as well, and ranges from 0 to 1. A negative correlation indicates that as one variable increases, the other variable decreases, and ranges from -1 to 0. Lastly, the r value of 0 indicates that there is no correlation (BYJU's, 2020).

Pearson's correlation coefficient formula is used to understand the correlation between variables:

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

r is the Pearson's correlation coefficient (what is being measured)

n is the total number of values

$\sum$ is the sum of the values

x is the values in the first data set

y is the values in the second data set

The correlation of determination, r^2, indicates the degree of strength between the x and the y variables (Dataz4s, 2020). This will calculate if the variables are closely correlated with reference to the degree.

Lastly, conditional probability determines the likelihood of an event occurring in accordance with another event (Glen, 2020). In this case, the conditional probability could be represented as follows:

"The probability that you will die from COVID-19, given you live in [state]."

The formula to find the conditional probability is as shown:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Event A is the probability that an individual lives in a specific state

Event B is the probability that an individual will die from COVID-19

Linear Regression:

In this investigation, the two variables that are studied are the average income per state and the COVID-19 deaths per state in the United States. It is crucial to have accurate data to conduct this investigation, so the Centers for Disease Control and Prevention (CDC) was used to determine the number of COVID-19 deaths in each state. Additionally, since the number of COVID-19 deaths is increasing daily, the number of COVID-19 fatalities were collected on October 29, 2020. In the data table, each of the 50 states' average income, total population, and

number of COVID-19 deaths were measured. In order to get an accurate representation of the deaths for each state, the number of deaths will need to be normalized per 100,000 people.

States	Average Income (PK, 2020)	2020 Population (World Population Review, 2020)	COVID-19 Deaths: October 29, 2020 (U.S. Department of Health & Human Services, 2020)	Number of deaths per 100,000 people
Alabama	$78,871.29	4,908,620	2,866	58.39
Alaska	$100,086.45	734,002	68	9.26
Arizona	$96,364.72	7,378,490	5,874	76.61
Arkansas	$77,637.36	3,039,000	1,812	59.62
California	$111,632.93	39,937,500	17,345	43.43
Colorado	$107,936.13	5,845,530	2,223	38.03
Connecticut	$117,303.22	3,563,080	4,577	128.46
Delaware	$102,639.68	982,895	685	69.69
District of Columbia	$134,385.32	720,687	642	89.08
Florida	$85,581.31	21,993,000	16,429	74.70
Georgia	$84,224.69	10,736,100	7,809	72.74
Hawaii	$105,978.87	1,412,690	211	14.94
Idaho	$91,721.86	1,826,160	573	31.38
Illinois	$103,958.50	12,659,700	9,775	77.21
Indiana	$87,139.09	6,745,350	4,130	61.23
Iowa	$86,536.71	3,179,850	1,634	51.39
Kansas	$96,719.61	2,910,360	975	33.50
Kentucky	$87,474.11	4,499,690	1,407	31.27
Louisiana	$78,124.94	4,645,180	5,837	125.66
Maine	$84,312.62	1,345,790	146	10.85
Maryland	$125,053.40	6,083,120	4,099	67.38

Massachusetts	$127,460.73	6,976,600	9,848	141.16
Michigan	$92,073.46	10,045,000	7,522	74.88
Minnesota	$104,195.36	5,700,670	2,402	42.14
Mississippi	$65,648.61	2,989,260	3,263	109.16
Missouri	$83,396.09	6,169,270	2,810	45.55
Montana	$81,638.21	1,086,760	297	27.33
Nebraska	$93,878.49	1,952,570	596	30.52
Nevada	$92,457.31	3,139,660	1,748	55.67
New Hampshire	$114,680.66	1,371,250	473	34.49
New Jersey	$119,305.58	8,936,570	16,285	182.23
New Mexico	$71,531.93	2,096,640	967	46.12
New York	$105,571.94	19,440,500	33,172	170.63
North Carolina	$84,727.74	10,611,900	4,157	39.17
North Dakota	$90,647.13	761,723	461	60.52
Ohio	$90,396.68	11,747,700	5,206	44.32
Oklahoma	$84,974.15	3,954,820	1,256	31.76
Oregon	$107,795.68	4,301,090	653	15.18
Pennsylvania	$99,681.52	12,820,900	8,666	67.59
Rhode Island	$98,980.03	1,056,160	1,177	111.44
South Carolina	$83,649.63	5,210,100	3,802	72.97
South Dakota	$83,574.96	903,027	375	41.53
Tennessee	$81,911.63	6,879,580	3,131	45.51
Texas	$98,362.04	29,472,300	17,504	59.39
Utah	$112,799.70	3,282,120	572	17.43
Vermont	$95,683.34	628,061	58	9.23
Virginia	$114,127.44	8,626,210	3,581	41.51
Washington	$110,680.46	7,797,100	2,296	29.45

West Virginia	$72,857.68	1,778,070	423	23.79
Wisconsin	$90,695.26	5,851,750	1,813	30.98
Wyoming	$84,500.23	567,025	68	11.99

Table 1. Representation of the average income, population, COVID-19 deaths, and normalized deaths per 100,000 for each state.

Graphing the data will help to determine the relationship between the average income and the number of deaths through the obtained normalized values. To do this, a scatterplot and a line of best fit will be used to represent the relationship.

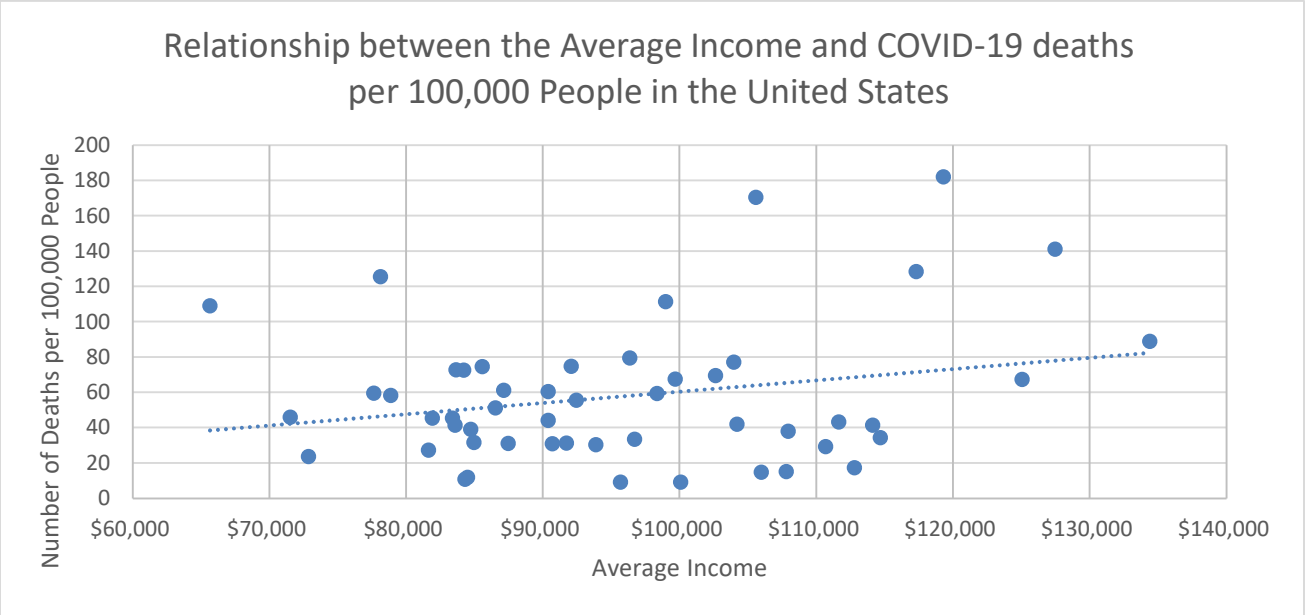

Figure 1. The relationship between the average income and COVID-19 deaths per 100,000 people in the United States as of October 29, 2020.

Figure 1 shows that there is a weak positive correlation between the average income and COVID-19 deaths per 100,000 people in the United States. Though there are a few outliers, it can be seen that the data mostly follow the trendline shown. The weak positive correlation suggests that the average income does not have a strong influence on the COVID-19 deaths in the United States. This, in turn, may emphasize that this virus can influence an individual of any income similarly.

To assess the relationship between average income and COVID-19 deaths, the slope and y-intercept need to be calculate using the linear regression line (WallStreetMojo, 2020). To find *a,* or the y-intercept, this formula should be used:

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

n= 51

$\sum x$= 4,881,316.03

$\sum y$= 2941.47

$(\sum x^2)$= 478,682,091,825.94

$\sum xy$= 288,859,327.09

$(\sum x)^2$= 23,829,691,298,659.60

After plugging in the desired number into the formula, it must be solved:

$$a = \frac{(2941.47)(478{,}682{,}091{,}825.94) - (4{,}881{,}316.03)(288{,}859{,}327.09)}{51(478{,}682{,}091{,}825.94) - (23{,}829{,}691{,}298{,}659.60)}$$

$$a = \frac{1.408 \times 10^{15} - 1.410 \times 10^{15}}{2.441 \times 10^{13} - 23{,}829{,}691{,}298{,}659.60}$$

$$a = \frac{-2 \times 10^{12}}{5.803 \times 10^{11}}$$

$$a = -3.45$$

Now, to complete the equation, *b* or the slope needs to be calculated. The equation is as follows:

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

$$b = \frac{51(288{,}859{,}327.09) - (4{,}881{,}316.03)(2941.47)}{51(478{,}682{,}091{,}825.94) - (23{,}829{,}691{,}298{,}659.60)}$$

$$b = \frac{1.473 \times 10^{10} - 1.436 \times 10^{10}}{2.441 \times 10^{13} - 23{,}829{,}691{,}298{,}659.60}$$

$$b = \frac{373{,}581{,}020}{5.803 \times 10^{11}}$$

$$b = 6.38 \times 10^{-4}$$

Therefore, the complete equation for the linear regression line is as follows:

$$y = -3.45 + 6.38 \times 10^{-4}x$$

This means that the line crosses the x axis at point -3.45. Additionally, this demonstrates that the if individuals made \$0 in income, there would be -3.45 COVID-19 deaths per 100,000 individuals. Furthermore, the slope of the data points is 6.38×10^{-4}. This shows that for each dollar increase in average income, there will be about a .0006 increase in COVID-19 deaths.

Pearson Correlation Coefficient:

To determine the exact correlation between the average income and the COVID-19 deaths in the United States, it is imperative to use the Pearson correlation coefficient formula as shown below:

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

n=51

$\sum xy$=288,859,327.09

$\sum x$= 4,881,316.03

$\sum y$= 2941.47

$(\sum x^2)$= 478,682,091,825.94

$(\sum x)^2$= 23,829,691,298,659.60

$(\sum y^2)$= 247,393.56

$(\sum y)^2$=8,652,262.73

$$r = \frac{51(288,859,327.09) - (4,881,316.03)(2941.47)}{\sqrt{51(478,682,091,825.94) - (23,829,691,298,659.60)}\sqrt{51(247,393.56) - (8,652,262.73)}}$$

$$r = \frac{1.473 \times 10^{16} - 1.436 \times 10^{10}}{\sqrt{2.441 \times 10^{13} - (23,829,691,298,659.60)}\sqrt{12,617,071.56 - (8,652,262.73)}}$$

$$r = \frac{373,581,020}{\sqrt{5.831 \times 10^{11}}\sqrt{3,964,808.83}}$$

$$r = \frac{373,581,020}{(323,405.76)(1,991.18)}$$

$$r = \frac{373,581,020}{1,520,480,752}$$

$$r = .2457$$

Therefore, the correlation of determination (r^2) can be calculated:

$$r^2 = (.2457)^2$$

$$r^2 = .0604$$

This means that the average income and COVID-19 deaths are not very closely correlated with each other. Though there is a .245 correlation, it is considered as a weak, positive correlation. Finding r^2 shows that there is a .0604 correlation of determination, which suggests that there is a very weak degree of strength between average income and COVID-19 deaths.

Conditional Probability:

Since I live in Florida, I would like to find the probability that my family or I may experience a fatality from COVID-19 for the specific state of Florida.

The data for the state of Florida is shown below:

State	Average Income	2020 Population	Number of COVID-19 deaths: October 29, 2020	Number of deaths per 100,000 people
Florida	$85,581.31	21,993,000	16,429	74.70

Table 2. Comparison between the average income, population, COVID-19 deaths, and normalized deaths per 100,000 for the state of Florida.

The conditional probability formula that will be utilized is represented is shown below:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

In this scenario, the formula could be restated as, "the probability that an individual will die from COVID-19, given they live in Florida." Event A is the probability of dying from COVID-19 and Event B is the probability an individual lives in Florida.

$$P(live\ in\ Florida\ /\ die\ from\ COVID-19) = \frac{P(die\ from\ COVID-19 \cap live\ in\ Florida)}{P(die\ from\ COVID-19)}$$

Additionally, the normalized data will need to be used in this formula.

$$P(live\ in\ Florida\ /\ die\ from\ COVID-19) = \frac{\frac{16{,}429}{21{,}993{,}000}}{\frac{21{,}993{,}000}{331{,}301{,}180}}$$

$$P(live\ in\ Florida\ /\ die\ from\ COVID-19) = \frac{.000747}{.066383}$$

$$P(live\ in\ Florida\ /\ die\ from\ COVID-19) = .0113$$

The equation shows that there is about a 1.13% chance that my family and I, all who live in Florida, could die from COVID-19. Though the number may be miniscule, it is important to encourage individuals to be safe and promote a healthy lifestyle, which includes wearing masks and social distancing.

Conclusion:

In order to investigate the relationship between average income and COVID-19 deaths in the United States, I calculated the linear regression line and the correlation coefficient. This allowed me to determine correlation and relationship between these two variables.

The calculations shown above illustrate that the average income and COVID-19 deaths have a weak correlation. As the number of COVID-19 deaths rise each day, it is almost the same likelihood that those with a significantly high income will have the same chance of dying from COVID-19 than those who have a low income. Therefore, it can be concluded that there is an equal chance that COVID-19 can kill a person with any amount of income.

The research that I conducted specifically applies to the United States because the average income was collected from each state in the United States. In addition, the COVID-19 deaths were calculated per each state, since the CDC provides statistics specific to the United States. However, the average income can be correlated with the number of COVID-19 deaths in countries outside of the United States to determine their relationship as well. In addition, the validity of a few of the sources may come into question, as they do not all come from experts. However, the CDC can qualify as a valid source, where much of the information about the COVID-19 deaths was found. In addition, the number of COVID-19 cases may be inaccurate as some deaths that are counted into the data may not be COVID-19 related. Furthermore, many states have prohibited the counting of COVID-19 cases and deaths, which can further hinder the data collected.

It is imperative to find these results as individuals need to take precautions no matter the income an individual receives. By finding this information, others may realize the importance of COVID-19 and the high possibility of transmission. Overall, COVID-19 can affect any individual in relation to their income, therefore it is crucial to take the right steps in maintaining a healthy and safe lifestyle.

References

BYJUS. “Pearson Correlation Formula- Pearson Correlation Interpretation: Byju's.” *BYJUS*, BYJU'S, 16 Sept. 2020, byjus.com/pearson-correlation-formula/.

Glen, Stephanie. “Conditional Probability: Definition & Examples.” *Statistics How To*, 7 Nov. 2020, www.statisticshowto.com/probability-and-statistics/statistics-definitions/conditional-probability-definition-examples/.

Glen, Stephanie. “Correlation Coefficient: Simple Definition, Formula, Easy Calculation Steps.” *Statistics How To*, 4 Dec. 2020, www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/.

PK. “Average Income by State, Median, Top & Percentiles [2020].” *DQYDJ*, 2 Nov. 2020, dqydj.com/average-income-by-state-median-top-percentiles/.

U.S. Department of Health & Human Services. “CDC COVID Data Tracker.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2020, covid.cdc.gov/covid-data-tracker/.

WallStreetMojo. “Regression Formula: Step by Step Calculation (with Examples).” *WallStreetMojo*, 7 Oct. 2020, www.wallstreetmojo.com/regression-formula/.

World Population Review. *US States - Ranked by Population 2020*, 2020, worldpopulationreview.com/states.

Yale. *Linear Regression*, www.stat.yale.edu/Courses/1997-98/101/linreg.htm.