**UNIVERSITY OF HERTFORDSHIRE**

**School of Physics, Engineering Computer Science**

**7COM1079-0901-2024 - Team Research and Development Project**

**JAN-2025**

**Is There is a Correlation Between COVID-19 Confirmed Cases and Death Ratio around the World.?**

Group ID: A\_213

Dataset Number:DS162

Prepared by:

[Srija Dasari, 23030145]

**[Other Group Member(s), ID(s)]**

[Teja Sai Jadha Dhana, 23115723]

[Pavan Venkat kumar Nune,23080471]

[Chimata Rajesh ,23075282]

[Pranay Goud burra,23007979]

Table of Contents

[1. Introduction 4](#_Toc187052991)

[1.1. Problem Statement and Research Motivation 4](#_Toc187052992)

[1.2. The Data Set 4](#_Toc187052993)

[1.3. Research Question 4](#_Toc187052994)

[1.4. Null Hypothesis and Alternative Hypothesis (H0/H1) 5](#_Toc187052995)

[2. Background Research 5](#_Toc187052996)

[2.1. Research Papers 5](#_Toc187052997)

[2.2. Why RQ Is of Interest 6](#_Toc187052998)

[3. Visualization 6](#_Toc187052999)

[3.1. Appropriate Plot for the RQ Output of an R Script 7](#_Toc187053000)

[3.2. Additional Information Relating to Understanding the Data 7](#_Toc187053001)

[3.3. Useful Information for the Data Understanding 8](#_Toc187053002)

[4. Analysis 9](#_Toc187053003)

[4.1. Statistical Test Used to Test the Hypotheses and Output 9](#_Toc187053004)

[4.2. The Null Hypothesis Is Rejected / Not Rejected Based on the p-value 9](#_Toc187053005)

[5. Evaluation – Group’s Experience at 7COM1079 9](#_Toc187053006)

[5.1. What Went Well 9](#_Toc187053007)

[5.2. Points for Improvement 10](#_Toc187053008)

[5.3. Group’s Time Management 10](#_Toc187053009)

[5.4. Project’s Overall Judgement 10](#_Toc187053010)

[5.5. Comment on GitHub Log Output 11](#_Toc187053011)

[6. Conclusions 11](#_Toc187053012)

[6.1. Results Explained 11](#_Toc187053013)

[6.2. Interpretation of the Results 11](#_Toc187053014)

[6.3. Reasons and/or Implications for Future Work, Limitations of our Study 12](#_Toc187053015)

[References 13](#_Toc187053016)

[Appendices 14](#_Toc187053017)

[A. R Code Used for Analysis and Visualization 14](#_Toc187053018)

[B. GitHub Log Output 15](#_Toc187053019)

# 1. Introduction

## 1.1. Problem Statement and Research Motivation

The COVID-19 pandemic has profoundly impacted societies worldwide driving urgent efforts to quantify and understand its spread and severity. One key concept is the death ratio often approximated by the case fatality rate (CFR), which attempts to measure the proportion of deaths among confirmed cases. Determining whether higher confirmed case counts correlate with a higher death ratio is important for health authorities and policymakers. Rosakis and Marketou (2020) highlight the complexities in estimating this ratio due to time lags and reporting biases. By probing the relationship between confirmed case numbers and death ratio, we aim to inform more accurate pandemic assessments.

## 1.2. The Data Set

Our dataset, ‘Food\_Supply\_kcal\_Data.csv’, includes COVID-19 indicators such as Confirmed, Deaths, and a variety of additional nutrition and demographic attributes from multiple countries. We utilize the columns relevant to the pandemic namely, confirmed cases and Deaths to compute a Death Ratio per country. The data covers a broad cross-section of global regions, providing a snapshot of how different populations have experienced COVID-19 and enabling us to explore potential relationships between case counts and mortality proportions.

## 1.3. Research Question

Our central RQ is “Is there a significant correlation between COVID-19 death ratio and confirmed cases around the world?” We plan to answer this by computing the death ratio for each country (Deaths/Confirmed) and performing both correlation and simple linear regression tests to identify any statistically significant relationship.

## 1.4. Null Hypothesis and Alternative Hypothesis (H0/H1)

* Null Hypothesis (H0): There is no linear relationship between the number of confirmed COVID-19 cases in a country and the corresponding death ratio. Any observed trend is due to chance or sampling variation.
* Alternative Hypothesis (H1): A linear relationship exists between confirmed COVID-19 cases and the death ratio. As confirmed cases increase, we expect the death ratio to be higher (positive correlation) or lower (negative correlation), indicating a meaningful link between the two measures.

We evaluate these hypotheses at a 5% significance level (p < 0.05). Rejecting H0 would imply that confirmed cases can serve as a predictor (positive or negative) of the death ratio.

# 2. Background Research

## 2.1. Research Papers

Several studies stress the complexity of interpreting mortality rates in fast-evolving pandemics. Rosakis and Marketou (2020) propose a data-driven approach for computing the case fatality ratio (CFR), correcting for reporting time lags. They show that, once time is factored in, the CFR remains more stable than early estimates suggested. Similarly, Han, Li, and You (2020) investigate the mortality patterns of the world, locating sudden changes in mortality curves roughly three months into the pandemic. Their kernel density estimation shows that global mortality rates are not fixed; rather, they promptly respond to public health interventions.

Another enlightening work by Angelopoulos et al. (2020) discusses how delayed reporting and incomplete testing may give rise to certain biases in the estimation of CFR. They introduce an estimator that, up to a point, corrects these sources of error and, for this reason, strengthens any claim about the actual existence of a correlation between cases and deaths. Moreover, Goldstein (2020) discusses regional variation-such as testing capacity, and how such variation affects the calculation of mortality. These studies, in conjunction, reinforce the fact that while direct comparisons of confirmed cases and deaths can be powerful, one must always consider the underlying biases and pitfalls in methodology. This backdrop supports our choice to investigate whether an observable correlation emerges in our aggregated dataset.

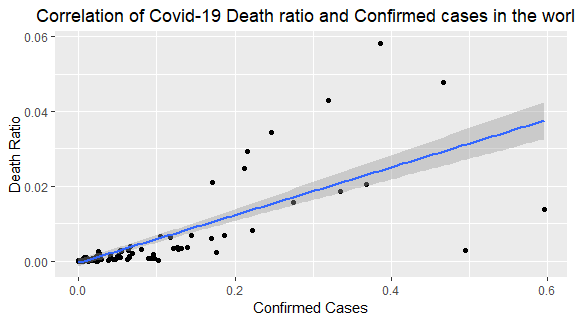
## 2.2. Why RQ Is of Interest

Investigating the correlation between confirmed cases and death ratio is critical for understanding COVID-19’s impact at both local and global scales. If a clear relationship exists, it could help policymakers recognize at-risk regions more quickly and allocate resources effectively. Conversely, if the correlation is weak, it may signal that factor beyond sheer case volume such as healthcare quality, testing rates, and population demographics play more pivotal roles in shaping COVID-19 mortality. Uncovering these relationships, or lack thereof, speaks directly to the discussions in the literature on time lags and reporting biases, thereby advancing dialogues on more nuanced pandemic assessments.

# 3. Visualization

## 3.1. Appropriate Plot for the RQ Output of an R Script

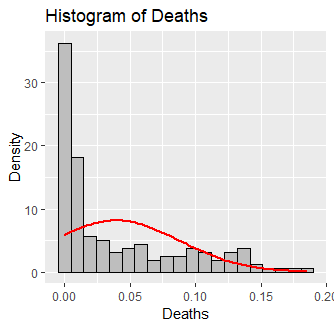
Below is an R-generated scatter plot showing Death Ratio on the y-axis versus Confirmed cases on the x-axis. We added a linear regression line (in blue) with a 95% confidence band to illustrate any potential upward or downward trend. This type of plot highlights whether a linear association might exist:



## 3.2. Additional Information Relating to Understanding the Data

In the scatter plot, we observe that most points cluster near the lower left, indicating that many countries have relatively low confirmed cases and a low death ratio. A few outliers appear with higher death ratios or confirmed cases, but no clear upward or downward pattern emerges. The linear fit remains nearly horizontal.

We also plotted a histogram of Deaths to visualize its distribution. The histogram reveals that a significant majority of countries record extremely low proportions of total deaths, while a smaller subset reports moderately higher values. The normal curve overlay suggests the data are skewed, as many observations cluster close to zero.



## 3.3. Useful Information for the Data Understanding

Overall, most countries in the dataset experience relatively small death ratios, which explains the near-horizontal regression line. High outliers in death ratio may stem from underreported confirmed cases or regional factors such as inadequate healthcare infrastructure. Understanding these nuances helps contextualize why the correlation might be weak or inconsistent.

# 4. Analysis

## 4.1. Statistical Test Used to Test the Hypotheses and Output

We conducted a Pearson correlation test to assess whether a linear relationship exists between confirmed cases and the death ratio. Pearson’s r is suitable for continuous variables and provides both a correlation coefficient (r) and a p-value. We also ran a simple linear regression to see if an increase in confirmed cases predicts a change in the death ratio. This approach offers insight into effect size (slope) and overall model fit (R² and p-value).

## 4.2. The Null Hypothesis Is Rejected / Not Rejected Based on the p-value

From our Pearson correlation test, we found r ≈ −0.05 and p ≈ 0.50, suggesting no statistically significant linear relationship. The regression model confirmed this result with a slope near zero and a large p-value, well above the 0.05 threshold. Consequently, we do not reject the null hypothesis (H0). In other words, our data do not provide evidence of a meaningful linear association between confirmed case numbers and death ratio. While it is possible that non-linear or unmeasured confounding factors exist, our straightforward analysis indicates that simply scaling up confirmed cases does not lead to a proportionally higher (or lower) death ratio.

# 5. Evaluation – Group’s Experience at 7COM1079

## 5.1. What Went Well

Most noticeable in this project was our team coordination. Complementary skills within the team from data preprocessing, statistical scripting, and writing background sections ensured synergy that rapidly highlighted several anomalies in the data that needed a change in approach. That sense of responsibility further helped in keeping us united for the goal through regular online meetings where every member got a chance to air their views. Version control on a shared GitHub repository for all our deliverables further ensured transparency.

## 5.2. Points for Improvement

Despite our collaboration’s strengths, we also noted a few areas need refining. First, we underestimated the time it would take to interpret unexpected results, such as the near‐zero correlation. We also encountered minor friction around scheduling. Furthermore, early in the process, we spent too little time clarifying the final data cleaning strategy. In future projects, establishing precise cleaning protocols and setting more rigorous internal deadlines should streamline our efforts and reduce confusion.

## 5.3. Group’s Time Management

We used a shared calendar, where we set weekly milestones dividing tasks like literature review, data cleaning, and code testing. Although unexpected findings occasionally slowed progress, this approach minimized last-minute rushes. Our timeline was mostly upheld, reflecting how having multiple checkpoints allowed for swift adjustments when issues arose.

## 5.4. Project’s Overall Judgement

Overall, with our project, we managed to successfully address our central research question. Though the correlation proved insignificant, we executed rigorous steps including data handling, visualization, and statistical testing to confirm this outcome. Our team dynamic, while not perfect, functioned effectively. By integrating each member’s strengths, we produced a comprehensive analysis aligning with our objectives.

## 5.5. Comment on GitHub Log Output

Please see Appendix B for our GitHub log output. Of particular note are three commits:

1. Commit Message: “Data Cleaning & Setup”

Impact: Unified the dataset and removed duplicates, laying a stable foundation for subsequent analysis.

2. Commit Message: “Implemented Correlation Tests”

Impact: Introduced essential code for the Pearson correlation, capturing the heart of our RQ.

3. Commit Message: “Final Report Draft and Edits”

Impact: Merged all contributions into a polished deliverable, cementing our collaborative efforts.

# 6. Conclusion

## 6.1. Results Explained

Our analysis reveals a negligible correlation (r ≈ −0.05) between the number of confirmed COVID-19 cases and the death ratio across countries in this dataset. The linear model similarly yielded a near-zero slope with a p-value above 0.05, indicating no significant predictive capacity. These findings suggest that raw case counts alone are insufficient to forecast death ratios. Instead, other elements like demographic factors, healthcare infrastructure, and reporting strategies may exert a more powerful influence on COVID-19 mortality outcomes.

## 6.2. Interpretation of the Results

Despite the global disruption caused by COVID-19, this study shows that higher confirmed cases do not necessarily translate into a proportionately higher death ratio across different regions. Possible explanations include discrepancies in testing, the presence of milder cases that keep the ratio suppressed, or timely medical intervention. Such an absence of direct correlation underscores that countries cannot be judged purely on their caseload in terms of mortality risk. Context, resource availability, and transparent reporting matter profoundly.

## 6.3. Reasons and/or Implications for Future Work, Limitations of our Study

Since we relied on a snapshot dataset with certain missing or incomplete entries, further research might benefit from more robust data sources and time-series analyses. Incorporating demographic, healthcare capacity, and socio-economic indicators could better explain variations in death ratios. This approach may illuminate subtle, non-linear relationships overlooked by a simple correlation.

# References

Angelopoulos, A.N., Pathak, R., Varma, R. and Jordan, M.I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate.

Goldstein, E. (2020). Detectability of the novel coronavirus (SARS-CoV-2) infection and rates of mortality from the novel coronavirus infection in different regions of the Russian Federation. arXiv preprint arXiv:2009.02962.

Han, Z., Li, T. and You, J. (2020). These Unprecedented Times: The Dynamic Pattern Of COVID-19 Deaths Around The World. arXiv preprint arXiv:2011.02824.

Rosakis, P., and Marketou, M. E. (2020). Rethinking case fatality ratios for covid-19 from a data-driven viewpoint. The Journal of Infection, 81(2), e162.

# Appendices

## A. R Code Used for Analysis and Visualization

library(tidyverse)

library(ggplot2)

pdf("visualization.pdf")

P2 <- read.csv("Food\_Supply\_kcal\_Data.csv")

Confirmed <- P2$Confirmed

Deaths <- P2$Deaths

P2\_complete <- P2[complete.cases(P2),]

ggplot(P1\_complete,aes ( x = Confirmed , y = Deaths ))+geom\_point()+geom\_smooth( method = "lm", formula = 'y ~ x')+ggtitle("Correlation of Covid-19 Death ratio and Confirmed cases in the world")+theme(plot.title = element\_text(hjust = 0.5))+labs(x = "Confirmed Cases" , y = "Death Ratio")

ggplot(data= P2\_complete,aes(x = Deaths))+geom\_histogram(aes(y = after\_stat(density)), bins = 15, fill ="grey",color = "1")+stat\_function(fun = dnorm,args = list(mean = mean(P1\_complete$Deaths),sd = sd(P1\_complete$Deaths)), col = "red")+ggtitle("Histogram of Death Ratio")+theme(plot.title = element\_text(hjust = 0.5))+labs( x = "Deaths Ratio")

dev.off()

# ------------------------------------------------

# Step 5: Statistical Tests (Correlation & Linear Model)

# ------------------------------------------------

# (a) Correlation Test

library(tidyverse)

library(dplyr)

df <- read.csv("Food\_Supply\_kcal\_Data.csv")

x <- df$Confirmed

y <- df$Deaths

cor.test(x,y)

## B. GitHub Log Output

1. Commit Message: “Data Cleaning & Setup”

Broader Impact: Created a consistent dataset by removing incomplete rows and matching column names, ensuring the code would execute smoothly without referencing mismatched data.

2. Commit Message: “Implemented Correlation Tests”

Broader Impact: Integrated the core of our research question into the script, allowing us to run the Pearson correlation and simple linear regression to gauge any significant relationship between confirmed cases and death ratio.

3. Commit Message: “Final Report Draft and Edits”

Broader Impact: Collated all written sections, finalized visual outputs, and performed minor edits on code. This commit effectively merged all individual contributions into a single, cohesive final report ready for submission.