Northeastern University
Mechanical and Industrial Engineering Department
IE 6200: Engineering Probability and Statistics
Prof. Rehab Ali
Fall 2023

"Global Influence of Nutritional Patterns on COVID-19"

Final Report

Group 6

Shalini Dutta, Venkata Kishan Madhav Grandhi, Siddhi Yeshwant Sonwalkar
Navisha Shetty, Jonna Jaidhitya

Disclaimer

The data that we have utilized for this study was compiled from publicly available sources including the Food and Agriculture Organization of the United Nations, the Population Reference Bureau, the Johns Hopkins Center for Systems Science and Engineering, and ChooseMyPlate.gov. While we endeavor to ensure the accuracy of the data, we make no justifications or representations of any kind, express or implied, about the completeness, accuracy, reliability, or suitability of the data for any purpose. This data and the inferences drawn are provided for educational and informational purposes only and should not be elucidated as professional advice or a substitute for consultation with qualified public health experts. The analysis and conclusions drawn are solely those of the author/s. **Any medical information is general and should not be used to diagnose or treat a health condition without consulting a qualified healthcare professional.**

Introduction

This project investigates the link between diet patterns and COVID-19 outcomes on a global scale. By analyzing international data on food supplies, obesity/hunger levels, and COVID-19 cases/deaths per country, the goal is to identify correlations that can inform dietary recommendations for building immunity and resilience against future pandemics. The study aims to guide research on lifestyle risk factors for infectious diseases, offering valuable insights into the role of healthy diets in combating diseases like COVID-19 with less severity.

Problem Statement

Our project investigates the impact of diets in 170 countries on COVID-19 recovery, analyzing data from 2020 to confirm associations between specific food groups and positive or negative outcomes.

We will employ descriptive statistics and hypothesis testing to analyze the data, focusing on key nutritional factors with the aim of recommending a healthy diet for COVID-19 prevention and recovery. However, challenges such as confounding variables (external factors), ethical concerns, and the dynamic nature of the virus complicate our research efforts.

The complex nature of the issue, influenced by genetics, vaccination, hygiene, and healthcare access, makes it challenging to establish direct links between diet and outcomes. Legal and regulatory issues, evolving scientific understanding, and ethical concerns further complicate the research. Results may also be affected by socioeconomic factors, healthcare access, and vaccination status.

In summary, while we aim to find the healthiest diet for COVID-19, the multifaceted nature of the problem and various limitations, including ethical and practical challenges, need careful consideration in drawing conclusions.

Project Goals

The project aims to explore how different diets around the world impact people's immunity to COVID-19. By analyzing the eating habits of 170 countries, we want to establish a connection between food choices and COVID-19 impacts.

Our research hypothesis suggests that the amount of animal products, meat, cereals, vegetables, and sugar consumed might be linked to the frequency and severity of COVID-19 cases. Conversely, the null hypothesis states that there's no meaningful correlation between these food groups and COVID-19.

In the first two weeks, we'll filter and classify the data, followed by data processing. The next two weeks involve using descriptive statistics and data visualization techniques like scatter plots and line charts to gain insights into dietary patterns. Moving to weeks five and six, we'll conduct regression, ANOVA, and hypothesis testing to determine if our hypothesis holds true. This rigorous statistical analysis will help identify a suitable diet for building immunity against COVID-19.

The overall goal is to contribute valuable information to nutritional departments worldwide. If successful, this project could guide dietary guidelines and public health policies to better prepare communities against future pandemics. The entire project will take two months, with progress updates every two weeks.

Data Collection

The project analyzes data from **170 countries** to understand how diets relate to COVID-19. We assume that countries with lower population densities might have fewer COVID-19 cases due to less frequent human contact in less crowded areas.

We acknowledge that factors like a country's genetic history, literacy rate, GDP, and healthcare infrastructure also affect COVID-19 transmission, though these are not in our dataset.

Data is collected from various sources, including the **Food and Agriculture Organization, Population Reference Bureau, Johns Hopkins Center for Systems Science and Engineering, and the Center for Nutrition Policy and Promotion.**

Our dataset covers 170 countries, categorizing food types and macronutrients like alcohol, animal products, cereals, fruits, meat, milk, starchy roots, sugar, sweeteners, vegetables, and more. We'll analyze how these factors relate to COVID-19 cases, deaths, recoveries, obesity, and undernourishment.

The project uses one of four tables present in the original dataset called - **Food_Supply_Quantity_kg.csv**. This table provides insights into food group intake percentages for each country.

Key variables include COVID-19-related percentages (**confirmed, death, recovered**), population, and **food intake percentages** for specific food groups.

Data Preparation –

The table extracted from Kaggle have been imported, cleaned and preprocessed as follows -

```
8  #IMPORT DATASETS
9  foodSupply <- read.csv("Food_Supply_Quantity_kg_Data.csv")
```

Fig. 1 – Importing Dataset into R

1. Handling Missing Values

```
33 #HANDLING MISSING VALUES
34 #1.To find total number of missing values in dataset
35 sum(is.na(foodSupply))
```

Fig. 2 – Finding missing values

```
> #HANDLING MISSING VALUES
> #1.To find total number of missing values in dataset
> sum(is.na(foodSupply))
[1] 36
```

Fig. 3 Number of missing values

```
46 #3. Removing rows with missing values
47 foodSupply2 <- foodSupply[complete.cases(foodSupply), ]
```

Fig. 4 Removing rows with missing values

2. Removing duplicate data –

```
58 #5. Remove duplicate data (can skip this since no duplicate data in datasets)
59 foodSupply2 <- distinct(foodSupply2)
```

Fig. 5 Removing rows with duplicate data

Our dataset is now prepared to start the visualization and correlate between variables to understand their dependencies and patterns.

Data Visualization

This project utilizes line graphs and scatter plots to evaluate preliminary data on potential dietary and lifestyle contributors to COVID-19 mortality risk during the 2020 pandemic.

Scatter Plots

The graphs depict data from the top 40 countries with the highest COVID-19 recovery rates among the 154 countries in our dataset after cleaning and pre-processing.

1. Animal Products

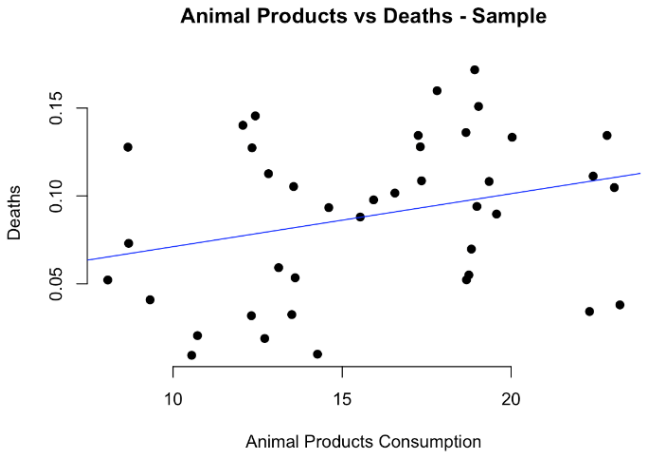


Fig. 6 Animal Product Consumption v/s Deaths

The scatter plot shows a weak positive correlation between COVID-19 death percentages and the consumption of animal products. As animal product consumption increases (X-axis), death percentages (Y-axis) also rise, with noticeable outliers and an even trend line spread. In summary, higher animal product consumption moderately correlates with elevated COVID-19 death rates with a **regression value around 0.28**, suggesting a potential impact on fatality.

2. Vegetables

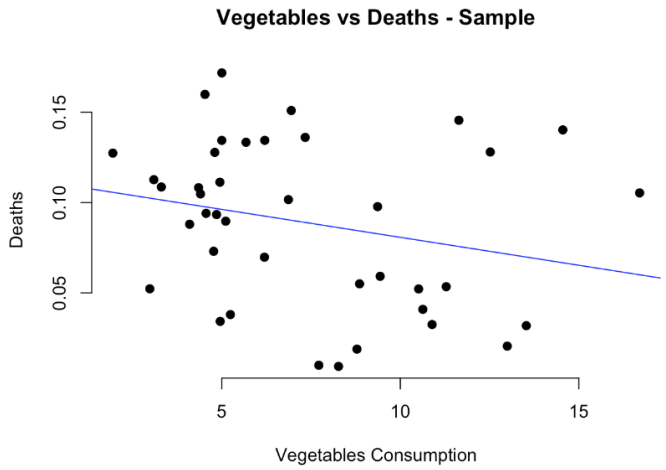


Fig. 7 Vegetables v/s Deaths

The scatter plot indicates a moderately negative correlation with **a regression value around 0.24**, between COVID-19 death percentages and vegetable consumption. As vegetable consumption increases, there is a corresponding decrease in death percentages, though outliers are present. In summary, this suggests a potential benefit, as countries with higher vegetable-rich diets tended to have somewhat lower COVID-19 death rates.

3. Meat

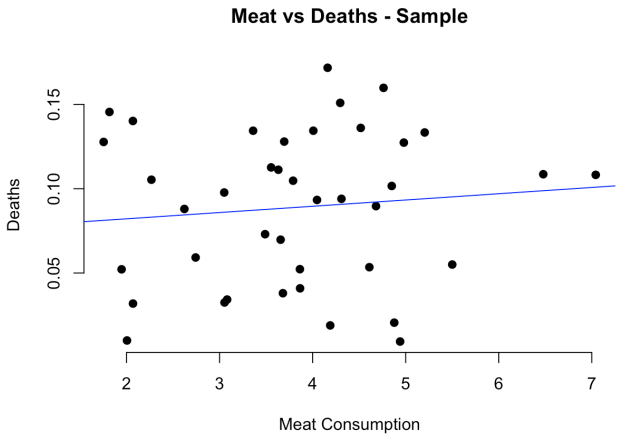


Fig. 8 Meat v/s Deaths

The scatter plot shows a moderate positive correlation with a **regression value around 0.1** between COVID-19 death percentages and the consumption of meat. As meat consumption increases (X-axis), death percentages (Y-axis) also rise, with noticeable outliers and a noticeable spread in the scatter. In summary, higher meat consumption moderately correlates with elevated COVID-19 death rates, suggesting a potential impact on fatality.

Line Charts

1. Sugars and sweeteners

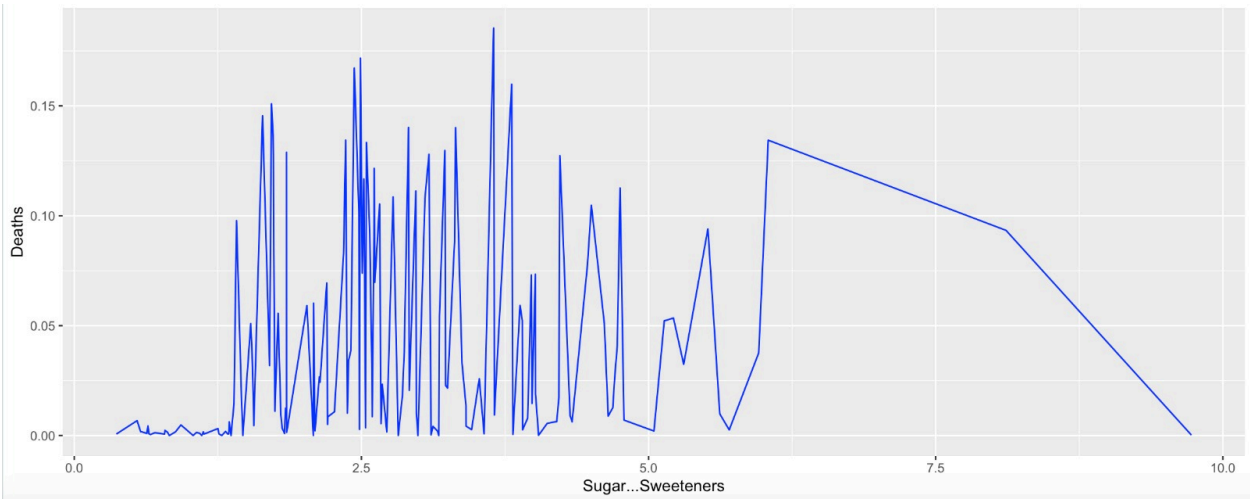


Fig. 9 Sugar v/s Deaths

The graph shows the connection between how much of people's diet comes from sugar and sweeteners (x-axis) and the percentage of COVID-19 deaths (y-axis) in 2020 across different

countries. Belgium had the highest mortality at 0.18%, linked to a 3.65% sugar/sweetener intake. The graph suggests a positive link, with a **regression value of 0.117**, meaning COVID-19 deaths rose with higher sugar/sweetener intake. This initial data suggests a possible connection, needing further study for confirmation. If proven, it implies higher sugar/sweetener diets might increase COVID-19 mortality risk.

2. Cereals

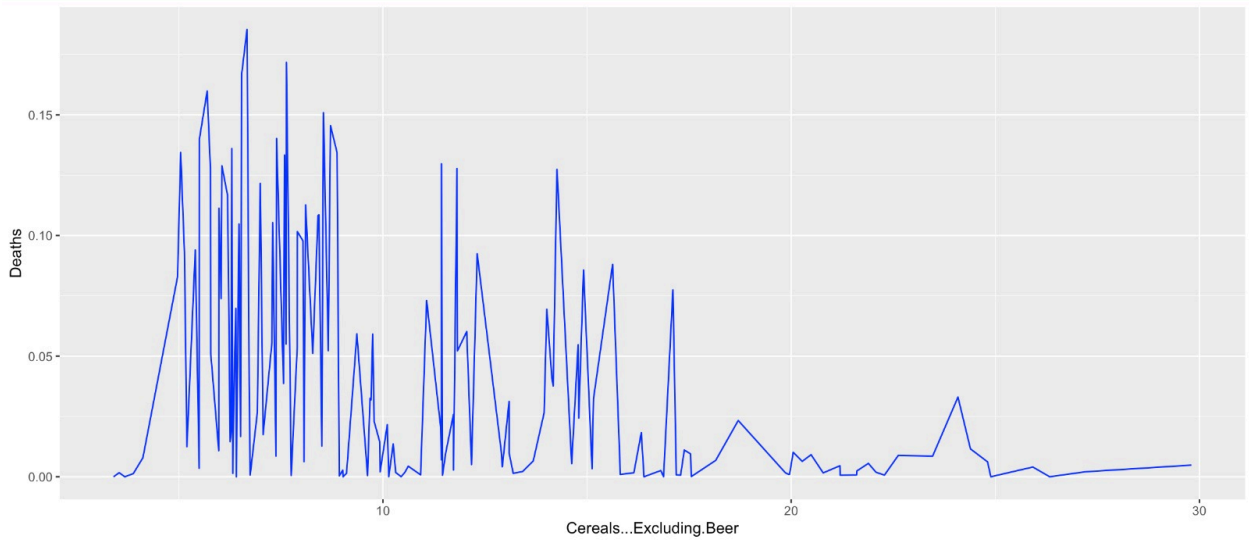


Fig. 9 Cereals v/s Deaths

The graph shows how the percentage of people's diet made up of cereals (excluding beer) relates to COVID-19 deaths in 2020 across different countries. As the cereal intake percentage goes up (x-axis), the COVID-19 death percentage goes down (y-axis). This suggests that countries with lower cereal consumption had higher COVID-19 death rates. Belgium, with only 6.6% of its diet from cereals, had the highest death rate at 0.18%. The negative correlation with a **regression value of 0.31** indicates that more cereal intake was linked to lower COVID-19 mortality risk. Further analysis is needed to confirm this and understand if cereal consumption indeed protects against COVID-19 deaths.

Statistical Analysis

I. Confidence Interval of Means

One-Sample T: Animal.Products, Meat, Sugar...Sweeteners, Vegetables, Cereals

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean	95% CI for μ
Animal.Products	20	17.544	3.317	0.742	(15.991, 19.096)
Meat	20	4.141	1.260	0.282	(3.551, 4.730)
Sugar...Sweeteners	20	3.473	1.665	0.372	(2.694, 4.253)
Vegetables	20	6.574	3.948	0.883	(4.727, 8.422)
Cereals...Excluding.Beer	20	7.669	2.748	0.614	(6.383, 8.955)

μ: mean of Animal.Products, Meat, Sugar...Sweeteners, Vegetables, Cereals...Excluding.Beer

Fig. 10 One-sample T-test - Descriptive statistics and Confidence Intervals

For this we have considered a sample of 20 countries with the highest recovery rates.

Animal Products:

In a holistic examination of 154 countries, the collective consumption of animal products is found to range from **15.991% to 19.096%.** Utilizing a 95% confidence interval with a sample mean of 17.544%, a standard deviation of 3.317%, and a sample size of 20 (t = 2.093, degrees of freedom = 19), the estimated average animal product consumption falls between 15.99% and 19.096%.

Meat:

In a holistic examination of 154 countries, the collective consumption of animal products is found to range from **3.551 % to 4.73%.** Utilizing a 95% confidence interval with a sample mean of 4.141%, a standard deviation of 1.26%, and a sample size of 20 (t = 2.093, degrees of freedom = 19), the estimated average animal product consumption falls between 3.551% and 4.73%.

Sugar Sweeteners:

n a holistic examination of 154 countries, the collective consumption of animal products is found to range from **2.694% to 4.253%.** Utilizing a 95% confidence interval with a sample mean of 3.473%, a standard deviation of 1.665%, and a sample size of 20 (t = 2.093, degrees of freedom = 19), the estimated average animal product consumption falls between 2.694% and 4.253%.

Vegetables:

In a holistic examination of 154 countries, the collective consumption of animal products is found to range from **4.727% to 8.422%.** Utilizing a 95% confidence interval with a sample mean of 6.574%, a standard deviation of 3.948%, and a sample size of 20 (t = 2.093, degrees of freedom = 19), the estimated average animal product consumption falls between 4.727% and 8.422%.

Cereals…Excluding Beer:

In a holistic examination of 154 countries, the collective consumption of animal products is found to range from **6.383% to 8.955%.** Utilizing a 95% confidence interval with a sample mean of 7.669 %, a standard deviation of 2.748%, and a sample size of 20 (t = 2.093, degrees of freedom = 19), the estimated average animal product consumption falls between 6.383% and 8.955%.

II. Confidence of Difference of Means

For this we use a sample of 20 countries with the highest recovery rates, and another sample of 20 countries with the lowest recovery rates.

1. Vegetables:

WORKSHEET 5

Two-Sample T-Test and CI: Vegetables (most recovered), Vegetables (least rec)

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Vegetables (most recovered)	20	6.57	3.95	0.88
Vegetables (least rec)	20	5.43	3.69	0.82

Estimation for Difference

Difference	95% CI for Difference
1.14	(-1.31, 3.59)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
0.95	37	0.350

Fig. 11 Two-sample T test – Vegetable consumption(highest v/s lowest recovery rate)

In countries with the most recovery rates, people eat, on average, 6.57% vegetables, while in those with the least recovery rates, it's 5.43%. With a 95% confidence level, the estimated average difference in vegetable consumption falls between -1.31% and 3.59%. This suggests, with 95% confidence, that the true difference in vegetable consumption between countries with the most and least recovery rates falls within this range. This suggests a difference in vegetable consumption between countries with the most and least recovery rates, with higher recovery rate countries tending to consume more vegetables.

2. Meat:

WORKSHEET 5

Two-Sample T-Test and CI: Meat (most recovered countries), Meat (least recovered)

Method

μ_1: mean of Meat (most recovered countries)
μ_2: mean of Meat (least recovered)
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Meat (most recovered countries)	20	4.14	1.26	0.28
Meat (least recovered)	20	3.21	1.70	0.38

Estimation for Difference

Difference	95% CI for Difference
0.933	(-0.027, 1.892)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

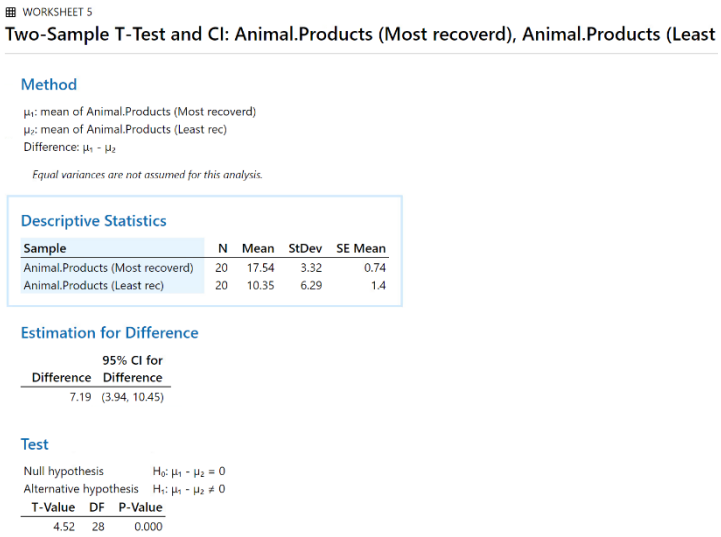
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
1.97	35	0.056

Fig. 11 Two-sample T test – Meat consumption(highest v/s lowest recovery rate)

In countries with the most recovery rates, people eat, on average, 4.14% meat, while in those with the least recovery rates, it's 3.21%. With a 95% confidence level, the estimated average difference in meat consumption falls between -0.027% and 1.892%. This suggests, with 95% confidence, that the true difference in meat consumption between countries with the most and least recovery rates falls within this range. From this, we can infer that there is a difference in meat consumption between countries with the most and least recovery rates. On average, those with higher recovery rates tend to consume more meat than those with lower recovery rates.

3. Animal Products:



WORKSHEET 5

Two-Sample T-Test and CI: Animal.Products (Most recoverd), Animal.Products (Least

Method

μ_1: mean of Animal.Products (Most recoverd)
μ_2: mean of Animal.Products (Least rec)
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Animal.Products (Most recoverd)	20	17.54	3.32	0.74
Animal.Products (Least rec)	20	10.35	6.29	1.4

Estimation for Difference

Difference	95% CI for Difference
7.19	(3.94, 10.45)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
4.52	28	0.000

Fig. 12 Two-sample T test – Animal Products consumption(highest v/s lowest recovery rate)

In countries with the most recovery rates, people consume, on average, 17.54% animal products, while in those with the least recovery rates, it's 10.35%. With a 95% confidence level, the estimated average difference in animal product consumption falls between 3.94% and 10.45%. This suggests, with 95% confidence, that the true difference in animal product consumption between countries with the most and least recovery rates falls within this range. From this, we can infer that there is a significant difference in animal product consumption between countries with the most and least recovery rates. On average, those with higher recovery rates tend to consume more animal products (17.54%) compared to those with lower recovery rates (10.35%).

4. Sugar:

WORKSHEET 5

Two-Sample T-Test and CI: Sugar...Sweeteners, Sugar...Sweeteners (least rec)

Method

μ_1: mean of Sugar...Sweeteners
μ_2: mean of Sugar...Sweeteners (least rec)
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sugar...Sweeteners	20	3.47	1.67	0.37
Sugar...Sweeteners (least rec)	20	2.59	2.09	0.47

Estimation for Difference

Difference	95% CI for Difference
0.885	(-0.327, 2.097)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
1.48	36	0.147

Fig. 13 Two-sample T test – Sugar consumption(highest v/s lowest recovery rate)

In countries with the most recovery rates, people consume, on average, 3.47% sugar and sweeteners, while in those with the least recovery rates, it's 2.59%. With a 95% confidence level, the estimated average difference in sugar and sweeteners consumption falls between -0.327% and 2.097%. This suggests, with 95% confidence, that the true difference in sugar and sweeteners consumption between countries with the most and least recovery rates falls within this range. From this, we can infer that there is a difference in sugar and sweeteners consumption between countries with the most and least recovery rates. On average, those with higher recovery rates tend to consume more sugar and sweeteners (3.47%) compared to those with lower recovery rates (2.59%).

5. Cereals:

WORKSHEET 5

Two-Sample T-Test and CI: Cereals (most recovered), Cereals (Least r

Method

μ_1: mean of Cereals (most recovered)
μ_2: mean of Cereals (Least rec)
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Cereals (most recovered)	20	7.67	2.75	0.61
Cereals (Least rec)	20	13.30	7.64	1.7

Estimation for Difference

Difference	95% CI for Difference
-5.63	(-9.39, -1.88)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-3.10	23	0.005

Fig. 14 Two-sample T test – Cereals consumption(highest v/s lowest recovery rate)

In countries with the most recovery rates, people consume, on average, 7.67% cereals, while in those with the least recovery rates, it's 13.30%. With a 95% confidence level, the estimated average difference in cereals consumption falls between -9.39% and -1.88%. This suggests, with 95% confidence, that the true difference in cereals consumption between countries with the most and least recovery rates falls within this range. From this, we can infer that there is a significant difference in cereals consumption between countries with the most and least recovery rates. On average, those with higher recovery rates tend to consume less cereals (7.67%) compared to those with lower recovery rates (13.30%).

III. Hypothesis Testing using Z-tests(Single-tailed)

1. Animal_Products

Population mean μ =12.1218

Population standard deviation σ = 6.0395

Sample size n = 40

Null Hypothesis H_0 :

The COVID-19 death rates are the same for individuals with different levels of animal products consumption in the population.

Alternative Hypothesis H_a :

If the consumption percentage of animal products exceeds the true mean consumption(>12.1218%) this would increase the death rates from COVID-19.

```
        One-sample z-Test

data:  foodSupply4$Animal.Products
z = 3.949, p-value = 3.924e-05
alternative hypothesis: true mean is greater than 12.12176
95 percent confidence interval:
 14.32205       NA
sample estimates:
mean of x
 15.89277
```

Since p<0.05, we reject the null hypothesis and accept the alternative hypothesis. Thus, an increased consumption of animal products decreases immunity in the population and reduces the chances of combatting COVID-19 and it's symptoms.

2. Vegetables

Population mean μ = 5.9715

Population standard deviation σ = 3.4915

Sample size n = 40

Null Hypothesis H_0 :

The COVID-19 death rates are the same for individuals with different levels of vegetable consumption in the population.

Alternative Hypothesis H_a :

If the consumption rate of vegetables exceeds the true mean consumption(>5.9715%), this would boost the immunity and recovery rate, thus reducing death rates.

```
        One-sample z-Test

data:  foodSupply4$Vegetables
z = 2.5369, p-value = 0.005592
alternative hypothesis: true mean is greater than 5.971525
95 percent confidence interval:
 6.463971       NA
sample estimates:
mean of x
 7.372018
```

Since p<0.05, we reject the null hypothesis and accept the alternative hypothesis. Thus, an increased consumption of vegetables increases immunity in the population and increases the chances of combatting COVID-19 and it's symptoms.

3. Cereals

Population mean μ = 12.0050

Population standard deviation σ = 5.9484

Sample size n = 40

Null Hypothesis H_0 :

The COVID-19 death rates are the same for individuals with different levels of cereal consumption in the population.

Alternative Hypothesis H_a :

If the consumption rate of cereals exceeds the true mean consumption(>12.0050), this would boost the immunity and recovery rate, thus reducing death rates.

```
        One-sample z-Test

data:  foodSupply4$Cereals...Excluding.Beer
z = -3.6672, p-value = 0.9999
alternative hypothesis: true mean is greater than 12.00502
95 percent confidence interval:
 7.008884       NA
sample estimates:
mean of x
 8.555922
```

Since p>0.05, we accept the null hypothesis and reject the alternative hypothesis. Thus, if the consumption of cereals is less than the true mean consumption of the population, then it doesn't affect the immunity/death rates from COVID-19 in any manner.

4. Meat

Population mean μ = 5.5697

Population standard deviation σ = 5.7786

Sample size n = 40

Null Hypothesis H_0 :

The COVID-19 death rates are the same for individuals with different levels of meat consumption in the population.

Alternative Hypothesis H_a :

If the consumption rate of meat exceeds the true mean consumption(>5.5697), this would adversely affect the immunity and cause an increase in death rates.

```
        One-sample z-Test

data:  foodSupply4$Meat
z = -1.9229, p-value = 0.9728
alternative hypothesis: true mean is greater than 5.569717
95 percent confidence interval:
 2.309917       NA
sample estimates:
mean of x
  3.81278
```

Since p>0.05, we accept the null hypothesis and reject the alternative hypothesis. Thus, if the consumption of meat is less than the true mean consumption of the population, it would positively impact the immunity and reduce fatalities from COVID-19.

5. Sugar

Population mean μ = 2.7324

Population standard deviation σ = 1.5119

Sample size n = 40

Null Hypothesis H_0 :

The COVID-19 death rates are the same for individuals with different levels of sugar consumption in the population.

Alternative Hypothesis H_a :

If the consumption rate of sugar exceeds the true mean consumption(>2.7324), this would positively affect the immunity and cause an decrease in death rates.

```
        One-sample z-Test

data:  foodSupply4$Sugar...Sweeteners
z = 2.4584, p-value = 0.006978
alternative hypothesis: true mean is greater than 2.732413
95 percent confidence interval:
 2.926887       NA
sample estimates:
mean of x
 3.320088
```

Since p<0.05, we reject the null hypothesis and accept the alternative hypothesis. Thus, an increased consumption of sugar positively affects immunity in the population and increases the chances of combatting COVID-19 and it's symptoms.

IV. ANOVA Test

ANOVA, or Analysis of Variance, is a statistical method that we have used to assess if there are any statistically significant differences between the means of the Death Rate, Recovery Rate and Confirmed Cases among the countries in the continents of Asia, Africa, Australia, Europe, North America, and South America. We have grouped the data based on the different continents and our analysis is as follows:

Hypothesis is common for all three categories.

Null Hypothesis H_0: Mean Death rate/ Recovery Rate / Confirmed Cases across all continents are equal.

Alternative Hypothesis H_a: Not all means across continents of Death Rate/ Recovery Rate/ Confirmed Cases are equal.

1. Death Rate

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	5	39.55	7.9096	70.21	0.000
Error	148	16.67	0.1127		
Total	153	56.22			

Means

Factor	N	Mean	StDev	95% CI
Africa	45	0.00864	0.01617	(-0.09024, 0.10751)
Asia	36	0.01968	0.02592	(-0.09087, 0.13022)
Australia	6	0.000709	0.001396	(-0.270063, 0.271481)
Europe	39	0.09205	0.04901	(-0.01415, 0.19826)
North America	17	0.0402	0.0484	(-0.1207, 0.2011)
South America	11	2.001	1.285	(1.801, 2.201)

Pooled StDev = 0.335634

As per the results of the ANOVA test for Death Rate, since the p value is significantly low (p<0.05), we reject the Null Hypothesis and accept the Alternative Hypothesis. We can infer that there is statistically significant differences in mean Death rate between at least two continents.

2. Recovery Rate

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	5	157.7	31.538	10.99	0.000
Error	148	424.6	2.869		
Total	153	582.3			

Means

Factor	N	Mean	StDev	95% CI
Africa	45	0.3547	0.5678	(-0.1443, 0.8537)
Asia	36	1.467	1.844	(0.910, 2.025)
Australia	6	0.0258	0.0408	(-1.3407, 1.3922)
Europe	39	2.942	2.454	(2.406, 3.478)
North America	17	1.168	1.695	(0.356, 1.980)
South America	11	2.001	1.285	(0.992, 3.010)

Pooled StDev = 1.69379

As per the results of the ANOVA test for Recovery Rate, since the p value is significantly low ($p<0.05$), we reject the Null Hypothesis and accept the Alternative Hypothesis. We can infer that there is statistically significant differences in mean Recovery rate between at least two continents.

3. **Confirmed Cases**

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	5	419.9	83.988	27.09	0.000
Error	148	458.8	3.100		
Total	153	878.7			

Means

Factor	N	Mean	StDev	95% CI
Africa	45	0.4169	0.6280	(-0.1017, 0.9356)
Asia	36	1.618	2.023	(1.038, 2.198)
Australia	6	0.0281	0.0448	(-1.3923, 1.4485)
Europe	39	4.671	2.191	(4.114, 5.228)
North America	17	1.937	2.462	(1.093, 2.781)
South America	11	2.271	1.370	(1.222, 3.320)

Pooled StDev = 1.76068

As per the results of the ANOVA test for Death Rate, since the p value is significantly low ($p<0.05$), we reject the Null Hypothesis and accept the Alternative Hypothesis. We can infer that there is statistically significant differences in mean Death rate between at least two continents.

Results and Conclusion

1. **Data Visualization:**

Scatter plots showed a moderate positive correlation between COVID-19 deaths and animal product/meat consumption, suggesting these may increase mortality risk.

Vegetables showed a negative correlation with deaths, indicating potential benefits.

Line graphs indicated a positive link between sugar/sweeteners and COVID-19 deaths, implying higher intake could increase risk.

Cereals had a negative correlation with deaths, suggesting potential protection.

2. **Confidence Intervals of Means:**

Confidence intervals were calculated to estimate the average consumption percentage ranges for animal products, meat, sugar and sweeteners, vegetables, and cereals for countries with the highest COVID-19 recovery rates.

3. **Confidence of Difference of Means:**

Comparing countries with the most and least recovery rates, 95% confidence intervals showed positive differences in consumption for vegetables, meat, animal products, and sugar for the higher recovery countries. Cereals showed lower consumption for higher recovery countries.

4. **Hypothesis Testing:**

Hypothesis tests found that increased consumption of animal products and sugar above the population average significantly increased COVID-19 death risk.

Increased vegetable consumption above the population average lowered COVID-19 death risk.

Meat and cereal consumption below their respective population averages showed no significant impact on COVID-19 immunity or deaths.

5. **ANOVA Test:**

ANOVA tests found statistically significant differences in mean COVID-19 death rate, recovery rate and confirmed cases between different continents.

In summary, the analysis suggests potential benefits of diets higher in vegetables and lower in animal products/sugar for resilience against COVID-19, while higher meat and cereals showed mixed impacts.

Key limitations around these variables are discussed below. Further research could help strengthen evidence on optimal diets.

Limitations

Trying to find the best diet for COVID-19 has some challenges. Figuring out if food directly causes certain outcomes or just relates to them, thinking about long-term effects, and accounting for individual differences and the virus's changing nature are some hurdles. Variables like genetics, vaccinations, hygiene, and healthcare access make it hard to say food directly causes results. Studying COVID-19 patients or those at risk has ethical issues.

Different virus versions might react differently to diets, and we're still learning the long-term effects. Other factors like social status, healthcare access, and vaccinations can muddy research, and self-reported data might not be reliable. Legal and scientific changes can also affect what's considered an ideal diet, so researchers need to be careful and consider these issues.

Proposed Next Steps and Future Work

To enhance our project, we've identified key areas for future development:

1. **Refine Data Collection:** Implement more precise data collection methods at the individual or community level, potentially through surveys for deeper insights.

2. **Long-Term Analysis:** Extend the analysis over a more extended period, spanning pre- and post-infection phases for a comprehensive understanding.

3. **Incorporate Additional Variables:** Broaden analysis by including other influencing variables, such as healthcare capacity, socioeconomic status and co-morbidities.

4. **Intervention Testing:** Explore opportunities to empirically test interventions, like public health communications, to gain practical insights.

5. **Customize Strategies:** Tailor strategies to specific communities based on detailed findings.

In summary, improving data collection, extending analysis duration, considering additional variables, testing interventions, and customizing strategies could enhance the robustness and applicability of our findings.

Bibliography

[1] https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset/

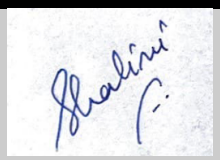
[2] Barrea, L., Grant, W. B., Frias-Toral, E., Vetrani, C., Verde, L., de Alteriis, G., ... & Muscogiuri, G. (2022). Dietary recommendations for post-COVID-19 syndrome. *Nutrients*, *14*(6), 1305.

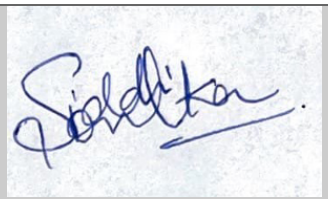
[3] Rodriguez-Leyva, D., & Pierce, G. N. (2021). The Impact of Nutrition on the COVID-19 Pandemic and the Impact of the COVID-19 Pandemic on Nutrition. *Nutrients*, *13*(6), 1752.

[4] https://www.kaggle.com/code/agnesa/eda-effect-of-diet-and-happiness-on-covid19

[5] https://drive.google.com/drive/folders/1xdKIqH-wztwpX76Wws2J2pT_fwl0BdwG?usp=drive_link – R-code, Minitab file, Dataset and other necessary files.

Collaboration

Shalini Dutta	**20%** • Formulated Project Goals and Problem Statement • Conducted data collection, finalized datasets from Kaggle • Performed data visualization, interpreted scatter plots • Applied two-tailed T-test, interpreted results and drew inferences • Utilized Z-test, formulated hypotheses and drew inferences • Developed R-code for scatter plots with regression line and data preparation for tests • Developed R-code for Z-test • Devised results and conclusions • Formatting, final proofreading	

Grandhi Venkata Kishan Madhav	**20%** • Defined project goals and problem statement • Collected and finalized Kaggle datasets • Interpreted and generated regression values from R-code scatter plots • Formulated null and alternative hypotheses for Z-test • Generated confidence intervals of means using 1-sample test in Minitab • Calculated confidence intervals of difference of means using 2-sample T-test in Minitab • Defined time-bound variables for project completion • Created frequency distribution table and histogram for food consumption items of different countries • Discussed and concluded project results • Formatting and proofreading the report	
Siddhi Yeshwant Sonwalkar	**20%** • Defined project goals and problem statement • Collected and finalized Kaggle datasets • Crafted project introduction • Conducted data visualization, interpreted line chart • Drew and documented inferences of confidence of mean • Drew and documented inferences of confidence of differences of mean • Created frequency distribution table and histogram for food consumption items of different countries • Documented future project enhancements • Discussed and concluded project results • Formatted and proofread the report	
Navisha Shetty	**20%** • Defined project goals and problem statement • Collected and finalized Kaggle datasets • Data Cleaning and Preprocessing in R	

	• Implemented scatter plots, line charts, etc. in R script for data visualization • Prepared data for ANOVA test • Segregated countries into different continents • Conducted ANOVA test using Minitab • Formulated hypotheses and drew inferences for ANOVA test • Discussed and concluded project results • Formatted and proofreading report	
Jonna Jaidhitya	**20%** • Defined project goals and problem statement • Collected and finalized Kaggle datasets • Performed data visualization, interpreted scatter plots • Documented the data collection process • Included project limitations • Discussed and documented the future enhancements • Concluded and presented project results • Conducted formatting and final proofreading