# R project

Sarfati Gabriel, de Thy Jean

2024-12-08

## Contents

*We solemnly declare that this report has been written by us without unauthorized external assistance, that it has not been submitted previously for evaluation, and that it has never been published in whole or in part. Any parts, groups of words, or ideas, no matter how limited, including tables, graphs, maps, etc., that are borrowed or reference other bibliographic sources are presented as such without exception.*

# 1 Introduction

On December 31, 2019, China first reported cases of pneumonia of unknown cause to the World Health Organization (WHO) in Wuhan, Hubei province. In about three months, the WHO declared SARS-CoV-2 (COVID-19) a global pandemic. Nations around the world started to close their external borders and impose nationwide lockdowns to prevent the spread of the disease. By December 31, 2020, the WHO reported 1,813,188 confirmed COVID-19 deaths worldwide. Furthermore, the pandemic caused a global economic downturn, with GDP contracting in most countries during 2020. Lockdowns and restrictions disrupted supply chains, reduced consumer spending, and caused widespread job losses, particularly in sectors like tourism, hospitality, and retail. Governments implemented massive fiscal and monetary stimulus measures to mitigate the impact, leading to increased public debt. The pandemic accelerated trends like remote work and digitization but exacerbated inequality, with low-income populations and developing countries hit hardest. Recovery has been uneven, with lingering effects like inflation, labour shortages, and altered global trade dynamics.

Throughout the project, our goal is to illustrate the impact of the COVID-19 pandemic on the economic development of countries worldwide between 2019 and 2020. More specifically, we want to evaluate the contraction by comparing countries' economic reliance on tourism, which has been a major threat to several developing nations.

Our analysis is based on a database published by the International Monetary Fund (IMF) in April 2024:

- Entire World Economic Outlook database
- To complete and compare these data, we used as well the resources provided by the World Bank in its World Development Indicators.

Our overall population is all countries (196 individuals) but is reduced to 179 since some countries did not disclose precisely their economic results. The analysis is based on the following variables, expressed in percentages of variations between the two periods :

- **General government gross debt variation (2019-2020)**: quantitative continue
- **Gross domestic product per capita in current price variations (2019-2020)**: quantitative continue
- **Variation of unemployment rate in % of total labour force (2019-2020)**: quantitative continue
- **Reliance of the economy on tourism (as a share of the nation's GDP)**: qualitative nominal
  - Low: $i < 5\%$
  - Medium: $5\% < i < 10\%$
  - High: $i > 10\%$
- **Continent**: qualitative nominal

### 1.1 Necessary set ups

```
rm(list=ls())
options(digits=3)
#First, we clean and set the environment

library(readxl) #--> charge and visualize the database
library(stringi) # --> enable R and Rmarkdown to read unusual characters
#We load the necessary libraries

#Charge the database in R
base <- read_excel("BDD_Stat.xlsx")
colnames(base) <- c("Countries", "ΔDebt", "ΔGDP", "ΔUnemployment", "Tourism",
  "Continent")

Total <- sum
All <- sum

base_clean <- na.omit(base)
```

When using AnyNA(base), we are informed that our database is not complete and lacks of a few information. The reason behind is that some countries are not disclosing their annual results. Henceforth, we must clean the database from any incomplete lines:

### 1.2 Sampling

```
# Setting a sample of 90 individuals out of 179 countries
set.seed(3)
sample_index <- sample(1:nrow(base_clean), size=90, replace=FALSE)
sample_data <- base_clean[sample_index,]
attach(sample_data)
```

While the 'attach' function might slow the running process of the code, it has greatly improved the coding process.

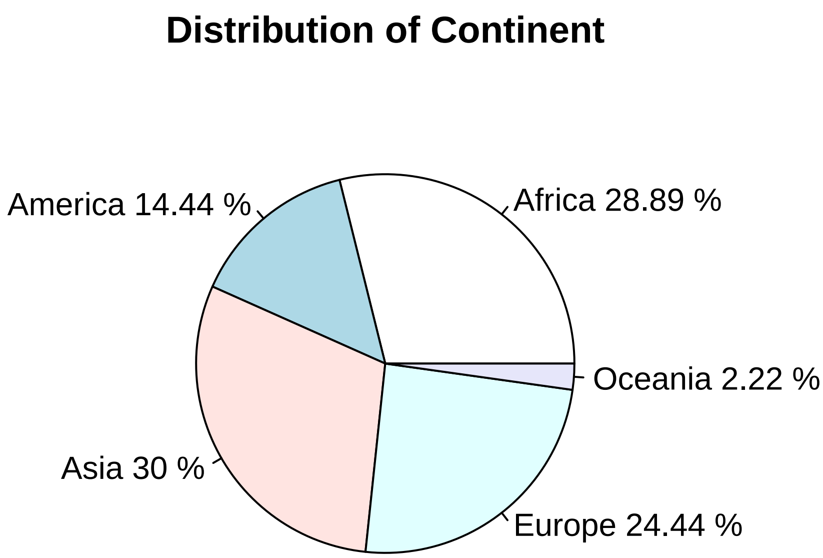
## 2 Analysis of the Frequencies (Graphs et Tables)

### 2.1 Distribution of Continents

```
addmargins(prop.table(table(Continent)),FUN=Total)
```

```
## Continent
##   Africa America    Asia  Europe Oceania   Total
##   0.2889  0.1444  0.3000  0.2444  0.0222  1.0000
```

```
tabC <- prop.table(table(Continent))*100
etiquetteC <- paste(rownames(tabC),round(tabC,2),'%')
pie(tabC,main='Distribution of Continent', labels=etiquetteC)
```



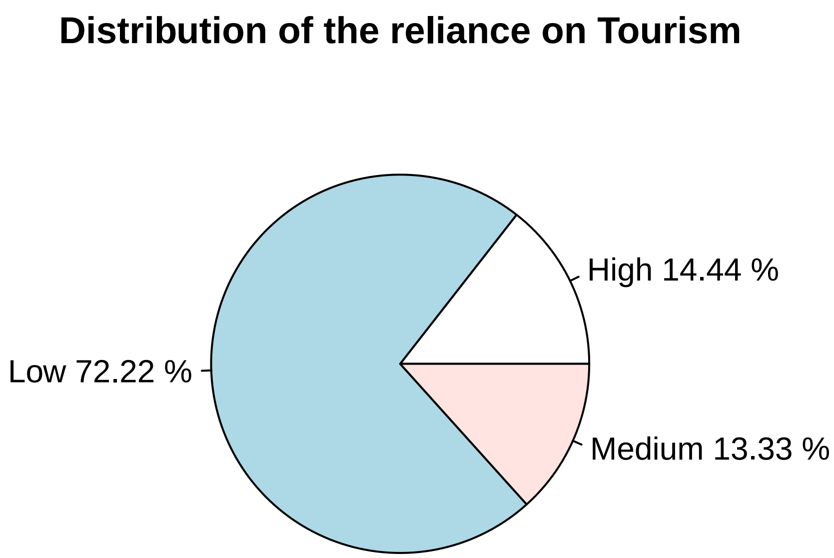

We can observe that the distribution of continents is coherent with reality, as Asian and African countries are more represented that others, specifically Oceania ones.

### 2.2 Distribution of countries' reliance on Tourism

```
addmargins(prop.table(table(Tourism)),FUN=Total)
```

```
## Tourism
##    High    Low Medium  Total
##   0.144  0.722  0.133  1.000
```

```
tabT <- prop.table(table(Tourism))*100
etiquetteT <- paste(rownames(tabT),round(tabT,2),'%')
pie(tabT,main='Distribution of the reliance on Tourism', labels=etiquetteT)
```



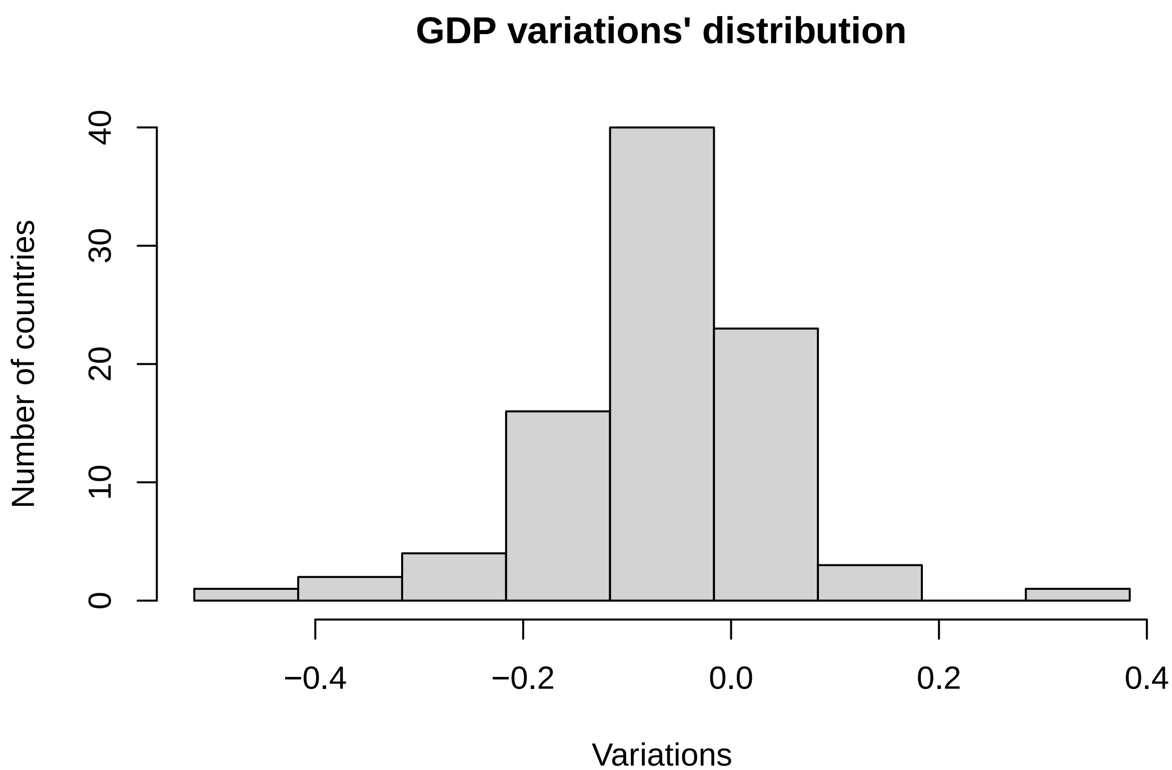

The pie chart showed that a great majority of countries have an overall GDP that is aggregated with less than 5% of the national tourism performance. Nonetheless, the distribution of countries that are highly dependent on tourism income has high extremes values that can influence the results.

### 2.3 GDP variations (ΔGDP)

```
breaks_GDP <- seq(min(ΔGDP),max(ΔGDP)+0.1,by=0.1)
GDP <- cut(ΔGDP,breaks_GDP)
addmargins(table(GDP),FUN=Total)
```

```
## GDP
##  (-0.516,-0.416]  (-0.416,-0.316]  (-0.316,-0.216]  (-0.216,-0.116]
##                0                2                4               16
## (-0.116,-0.0164] (-0.0164,0.0836]   (0.0836,0.184]    (0.184,0.284]
##               40               23                3                0
##    (0.284,0.384]            Total
##                1               89
```

```
hist(ΔGDP,breaks=breaks_GDP,
     main="GDP variations' distribution",
     xlab='Variations',
     ylab='Number of countries')
```




The distribution of the GDP performance between 2019 and 2020 seems bell-shaped with most of countries having a performance located between -10% and 0%.

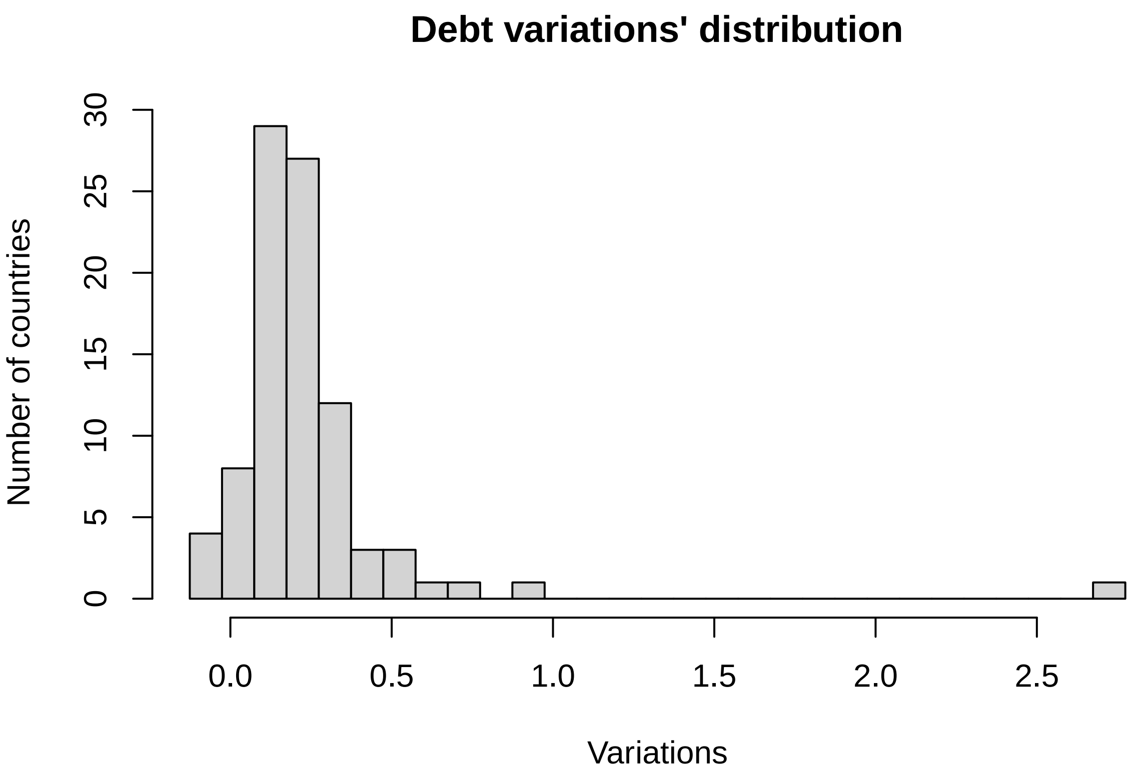
## 2.4 Debt variations (ΔDebt)

```
breaks_Debt <- seq(min(ΔDebt),max(ΔDebt)+0.1,by=0.1)
Debt <- cut(ΔDebt,breaks_Debt)
addmargins(table(Debt),FUN=Total)
```

```
## Debt
## (-0.126,-0.0259] (-0.0259,0.0741]   (0.0741,0.174]    (0.174,0.274]
##                3                8               29               27
##    (0.274,0.374]    (0.374,0.474]    (0.474,0.574]    (0.574,0.674]
##               12                3                3                1
##    (0.674,0.774]    (0.774,0.874]    (0.874,0.974]     (0.974,1.07]
##                1                0                1                0
##      (1.07,1.17]      (1.17,1.27]      (1.27,1.37]      (1.37,1.47]
##                0                0                0                0
##      (1.47,1.57]      (1.57,1.67]      (1.67,1.77]      (1.77,1.87]
##                0                0                0                0
##      (1.87,1.97]      (1.97,2.07]      (2.07,2.17]      (2.17,2.27]
##                0                0                0                0
##      (2.27,2.37]      (2.37,2.47]      (2.47,2.57]      (2.57,2.67]
##                0                0                0                0
```

```
##      (2.67,2.77]           Total
##                1              89
```

```
hist(ΔDebt,breaks=breaks_Debt,
     main="Debt variations' distribution",
     xlab='Variations',
     ylab='Number of countries')
```




Overall, a large majority of countries have expanded their national deficit during the period to support the slowdown in international markets and in the distribution chain with the main supplier.

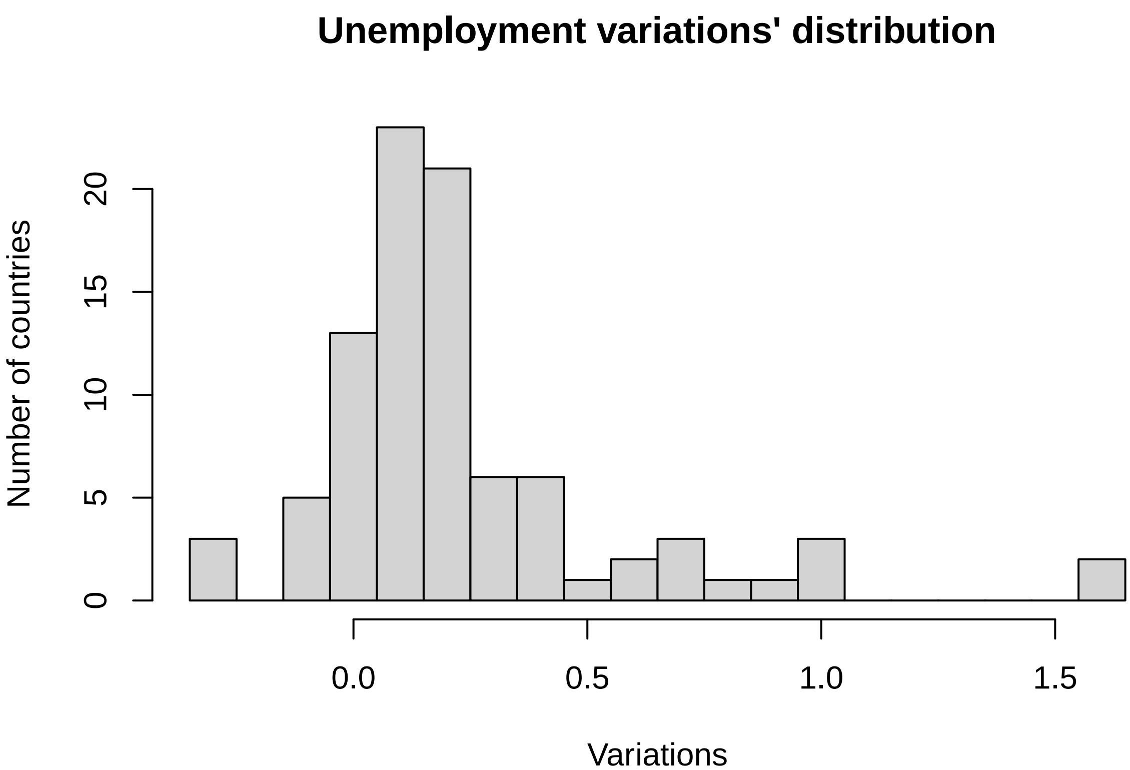
### 2.5 Unemployment variations ($\Delta$Unemployment)

```
breaks_Unemployment <- seq(min(ΔUnemployment),max(ΔUnemployment)+0.1,by=0.1)
Unemployment <- cut(ΔUnemployment,breaks_Unemployment)
addmargins(table(Unemployment),FUN=Total)
```

```
## Unemployment
## (-0.35,-0.25] (-0.25,-0.15] (-0.15,-0.05]  (-0.05,0.05]   (0.05,0.15]
##             2             0             5            13            23
##   (0.15,0.25]   (0.25,0.35]   (0.35,0.45]   (0.45,0.55]   (0.55,0.65]
##            20             7             6             1             2
##   (0.65,0.75]   (0.75,0.85]   (0.85,0.95]   (0.95,1.05]   (1.05,1.15]
##             3             1             1             3             0
```

```
## (1.15,1.25] (1.25,1.35] (1.35,1.45] (1.45,1.55] (1.55,1.65]
##           0           0           0           0           2
##       Total
##          89
```

```r
hist(ΔUnemployment,breaks=breaks_Unemployment,
     main="Unemployment variations' distribution",
     xlab='Variations',
     ylab='Number of countries')
```




The pandemic greatly affected countries that based most of their national income on tourism, especially countries in the pacific and islands (Fiji, Maldives, Aruba…)

Nevertheless, it is essential to note that most of these countries have smaller populations than Western developed countries. So when a significant proportion of the population (at scale) finds itself out of work, the figures explode.

# 3 Point estimation and Confidence interval for GDP's evolution

## 3.1 Point estimation

Here, since n > 30, the Central Limit Theorem tells us that the standardization of the empirical mean follows a normal distribution. The sample mean is used as a point estimation for the population. It is an unbiased and consistent estimator.

```r
n <- length(ΔGDP)

xbar_GDP <- mean(ΔGDP)
xbar_GDP
```

```
## [1] -0.0689
```

We want to compare our results with the mean of the GDP from the entire population:

```r
N <- nrow(base_clean)
X_G <- base_clean$ΔGDP
Xbar_Gpop <- mean(X_G)
Xbar_Gpop
```

```
## [1] -0.061
```

We can observe that for a sample of 90 individuals, our estimation of the mean is not far from the actual mean for the population.

### 3.2 Confidence interval estimation at 95%

Here we don't know the standard-deviation of the population. As we use a sample of 90 individuals, we can estimate the standard deviation of the population by the adjusted sample standard deviation.

To help us construct the interval, we will use quantiles of a Student Law of (n-1) degree of freedom.

We can therefore construct our two-sided confidence interval at 95% and 90% confidence level.

```r
scor_GDP <- sd(ΔGDP)

a1 <- xbar_GDP - qt(0.975, n-1) * scor_GDP / sqrt(n)
b1 <- xbar_GDP + qt(0.975, n-1) * scor_GDP / sqrt(n)
c(a1, b1)
```

```
## [1] -0.0922 -0.0457
```

We can affirm with a level of confidence of 95% that the average performance of the sample is located between -9.22% and -4.57%.

This shows that, overall, this has been a critical period for the development of economies around the world.

### 3.3 Confidence interval estimation at level 90%

```
a2 <- xbar_GDP - qt(0.95,n-1)*scor_GDP/sqrt(n)
b2 <- xbar_GDP + qt(0.95,n-1)*scor_GDP/sqrt(n)
c(a2,b2)
```

```
## [1] -0.0884 -0.0495
```

We assumed a lower level of confidence when determining the interval. Therefore, without touching the sample size, this increased the precision of the interval. It is logical since it can now be assumed that we are taking more risks to determine a narrower interval.

# 4 Conformity Test for Unemployement's evolution

```
n <- nrow(sample_data)
xbar_Unemployment <- mean(ΔUnemployment)

#Estimate of the average variation in the unemployment rate in our population
N <- nrow(base_clean)
X_U <- base_clean$ΔUnemployment
Xbar_Upop <- mean(X_U)
c(Xbar_Upop, xbar_Unemployment)
```

```
## [1] 0.243 0.233
```

We want to estimate the conformity of our estimation of the unemployment mean to the one of the population. We want to know if, statistically, the mean found in the sample is equal to the mean of the population. We found 24.3% percent of variation in average.

Would the sample be statistically equal at a significance level of 5%?

The standard deviation is unknown but we have a sample population greater than 30.

Hypothesis:

$$H_0 : m = 0.243$$
$$H_1 : m \neq 0.243$$

Test Statistics: sample mean

Decision rule: we will reject the hypothesis that the sample mean is statistically equal to the population mean if the critical probability is below 5%

```
t.test(ΔUnemployment,mu=Xbar_Upop,alternative='two.sided')
```

```
##
##  One Sample t-test
##
```

```
## data:  ΔUnemployment
## t = -0.3, df = 89, p-value = 0.8
## alternative hypothesis: true mean is not equal to 0.243
## 95 percent confidence interval:
##  0.162 0.304
## sample estimates:
## mean of x
##     0.233
```

We fail to reject that the sample mean is statistically equal to the population mean. This is because the probability of obtaining results as extreme or more extreme than those observed, under $H_0$, is 40%. This indicates that the data are consistent with what the null hypothesis predicts.

- At a level of significance equal to 10%, the conclusion is therefore the same.

# 5 Homogeneity Test

Problem: Does a country's dependence on tourism influence the impact of COVID on the country's economical growth in 2020?

## 5.1 Subgroups of the population

```r
group1 <- subset(sample_data,Tourism=='High')
group2 <- subset(sample_data,Tourism=='Low')
x1 <- group1$ΔGDP
x2 <- group2$ΔGDP
xbar1 <- mean(x1)
xbar2 <- mean(x2)
xbar1
```

```
## [1] -0.141
```

```r
xbar2
```

```
## [1] -0.0552
```

In this sample of 90 countries, the average variation in GDP in countries with a high dependence on tourism (xbar1 = -14.1%) appears to be much lower than in countries with a low dependence on tourism (xbar2 = -5.52%).

We will therefore test the hypothesis that the countries of the world least dependent on tourism have been less affected economically by covid than those most dependent on tourism.

### 5.2 Step 1 : Show that we can assume equal variances

```r
nrow(group1)
```

```
## [1] 13
```

```r
nrow(group2)
```

```
## [1] 65
```

We work with sub-populations for which we don't know the variances. Given that our sample comprises less than 30 individuals in one of the sub-populations (13 in group 1 and 65 in group 2), we must be able to assume with some confidence that the variances are equal.

We proceed to a two tailed test:

Hypothesis:

$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 \neq \sigma_2$$

```r
var.test(x1,x2,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.9, num df = 12, denom df = 64, p-value = 0.8
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.40 2.45
## sample estimates:
## ratio of variances
##              0.862
```

We cannot reject $H_0$ (equality of variances). Hence, we can assume at a 95% confidence that the variances of sub-populations 1 and 2 are equal.

### 5.3 Step 2 : T-test

Let $m_1$ be the sample mean of sub-population 1, that is countries with high dependency on tourism and $m_2$, the sample mean of sub-population 2, countries with low dependency on tourism.

Let's test the following hypothesis in a left-tailed test:

Hypothesis:

$$H_0 : m_1 = m_2$$

$$H1 : m_1 < m_2$$

```
t_test_result <- t.test(x1, x2, alternative='less',var.equal = TRUE)
print(t_test_result)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = -2, df = 76, p-value = 0.008
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf -0.0277
## sample estimates:
## mean of x mean of y
##   -0.1405   -0.0552
```

Here we can reject the equality of the mean ($H_0$) and assume $m_1 < m_2$ ($H_1$) at a confidence level of 95% as p-value is lower than 0.05

# 6 Test of Independence (Chi-square Test) : GDP and Tourism

We want to know if, economically, the pandemic had a greater negative impact on the variations of GDP performance in tourism-reliant countries after the collapse of the tourism market. Is GDP performance independent from the openness of a country to tourism?

Statistically, this would be translated through the observation of a correlation between the two variables.

## 6.1 Contingency table

```
GDP_borne <- c(-Inf, -0.05, Inf)
GDP_cut <- cut(ΔGDP, breaks = GDP_borne, right = TRUE)
TC_emp <- table(GDP_cut, Tourism)
addmargins(TC_emp,FUN=Total)
```

```
## Margins computed over dimensions
## in the following order:
## 1: GDP_cut
## 2: Tourism
```

```
##                Tourism
## GDP_cut         High Low Medium Total
##   (-Inf,-0.05]    11  33      5    49
##   (-0.05, Inf]     2  32      7    41
##   Total           13  65     12    90
```

We first have to build a table that compare the two variables.

## 6.2 Table of line profiles

```r
addmargins(prop.table(addmargins(TC_emp,1,FUN=All),1),2,FUN=Total)
```

```
##                Tourism
## GDP_cut           High    Low Medium  Total
##   (-Inf,-0.05] 0.2245 0.6735 0.1020 1.0000
##   (-0.05, Inf] 0.0488 0.7805 0.1707 1.0000
##   All          0.1444 0.7222 0.1333 1.0000
```

## 6.3 Table of column profiles

```r
addmargins(prop.table(addmargins(TC_emp,2),2),1)
```

```
##                Tourism
## GDP_cut          High   Low Medium   Sum
##   (-Inf,-0.05] 0.846 0.508  0.417 0.544
##   (-0.05, Inf] 0.154 0.492  0.583 0.456
##   Sum          1.000 1.000  1.000 1.000
```

From the contingency tables, it is hard to draw conclusions about the independence of the two variables.

## 6.4 Chi-square test of independence

We then need to conduct the test:

```r
result <- chisq.test(GDP_cut,Tourism, correct = FALSE)
result
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  GDP_cut and Tourism
## X-squared = 6, df = 2, p-value = 0.05
```

The p-value is equal to the level of significance of 5%. Therefore, we cannot conclude at this level of risk that the two variables are dependent.

Nonetheless, the hypothesis of independence is rejected at a level of significance of 10%.

We can conclude from an economical perspective that the GDP performance is impacted by the reliance of a country on tourism.

Let's finally verify to improve the accuracy of our test that the cross tabulations were all greater than 4. Otherwise, we would not be able to use the law:

Theoretical contingency table:

```
result$expected
```

```
##               Tourism
## GDP_cut         High  Low Medium
##   (-Inf,-0.05] 7.08 35.4   6.53
##   (-0.05, Inf] 5.92 29.6   5.47
```

# 7 Conclusion

This report explored the economic impact of the COVID-19 pandemic on countries with varying reliance on tourism between 2019 and 2020. Our statistical analysis revealed significant relationships between economic indicators such as GDP variation, debt expansion, and unemployment growth.

The analysis confirmed that countries heavily dependent on tourism experienced more severe GDP contractions. Through hypothesis testing, we found statistical evidence supporting that GDP decline was more significant in these nations compared to those less reliant on tourism. A t-test confirmed the negative effect, while a chi-square independence test showed that, at a 10% significance level, GDP variation was dependent on tourism reliance.

Additionally, public debt increased substantially across all continents, reflecting large-scale fiscal responses to the crisis. Unemployment spikes were more pronounced in small, tourism-dependent economies due to limited labor market flexibility.

Overall, our findings align with economic expectations: nations reliant on tourism faced greater economic disruptions due to border closures and reduced travel. These results highlight the vulnerability of tourism-based economies and underline the necessity for economic diversification and contingency planning in global crises.