06 HAZİRAN 2023

INE2002 TERM PROJECT

"Is there a relationship between life expectancy at birth and GNI per capita"

YAĞIZ EFE KOMİT-2002432
EREN BOZ-2104131
KEREM ÖZUTKU-2104368

TABLE OF CONTENTS

1. DATA COLLECTION PHASE

1.1 How Was Data Collected?

The dataset (World Human Development Report 2021) is taken from kaggle.com website. (The link of the dataset: https://www.kaggle.com/datasets/rajkumarpandey02/human-development-index-and-components?resource=download)

The contents that the dataset includes are Human Development Index (HDI), life expectancy at birth, expected years of schooling, mean years of schooling, gross national income (GNI). We focus on gross national income per capita and life expectancy at birth. Our purpose is to find out if there is a relationship between GNI per capita and life expectancy at birth. We focus on these two but also used mean years of schooling component. These data are for 2021.

1.2 Creating a Sample

The dataset contains 195 countries, this can be considered as a population. It is a csv file, for this reason it is easy to manage. First, we created a sample. Since it's not a statistical process and simple to use, we used Python language. The Python codes we used are in the figure below:

```python
import pandas as pd
# Upload the data file
data = pd.read_csv('/Users/yagizefekomit/Desktop/ine2002_project/humanindex.csv')

# Take a sample from the data, the size of the sample is n=35
sample_data = data.sample(n=35)

# Save the sample data to a new CSV file
sample_data.to_csv('new_sample.csv', index=False)  # new_sample.csv is our sample.
```

Figure 1.2.1: The Python codes to create a sample from data.

We have used random sampling method to avoid bias. Thus, we have obtained the sample completely at random. One of the difficulties we faced after creating a sample was that the type of the elements in one column was not numerical. We changed all of the elements in the column which are string type with integer by hand.

2. NORMAL DISTRIBUTION

Before to start processes in R, we the starters which are used in every R process have to be explained. Because we didn't put these codes to every figure but they're included.

```r
newSample <- read.csv("/Users/yagizefekomit/Desktop/ine2002_project/new_sample_2.csv")
leab <- newSample$Life.expectancy.at.birth
gni <- newSample$Gross.national.income..GNI..per.capita
schooling <- newSample$Mean.years.of.schooling
```

Figure 2.0.1: The starters used in every R process.

First, the csv file has been read and became a dataset which named '*newSample*'. The life expectancy at birth column stored as *'leab'*, GNI per capita stored as *'gni'* and mean years of schooling stored as *'schooling'.* These are used in every R process.

2.1 Checking for Normality

It is important to understand whether the data which is used is acceptable as a normal distribution or not. There are some signs and tests for this. Let's look at the histogram first. If it is normally distributed, histogram have to appear bell shaped. The function to plot the histogram is '*hist(leab)'*, '*leab'* stands for data which is going to be plotted.

Figure 2.1.1: The histogram of life expectancy at birth.

Looks approximately bell-shaped, which is a sign for normal distributed data. Also, there is another sign to support that the data is normal distribution is confidence intervals. We're going to observe confidence interval for mean and also interquartile range. But first, let's calculate the median and the mean in R.

```
> summary(leab)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.50   63.05   72.80   70.66   77.65   85.50
```

Figure 2.1.2: The summary of the data.

As stated in Figure 2.1.2, median is 72.80 and mean is 70.66. Now, let's look at The *Pearson index of skewness*. If the index is greater than or equal to +1 or less than or equal to -1, it can be concluded that the data is significantly skewed.

$$PC = \frac{3(X - median)}{s}$$

To apply this method in R, the codes that used are shown in the figure below.

```r
pearson_skewness <- function(x) {
  mean_val <- mean(x)
  median_val <- median(x)
  sd_val <- sd(x)

  skewness <- 3 * (mean_val - median_val) / sd_val
  return(skewness)
}

# Example usage with a numeric vector named 'data'
data <- leab
skewness_value <- pearson_skewness(data)
print(skewness_value)
```

Figure 2.1.3: The algorithm to calculate Pearson index.

The result of the equation is -0.7144. It is between -1 and 1. Therefore, Pearson index also supports the normality.

2.2 Confidence Intervals

A confidence interval is a range of values that is calculated based on sample data and a specified level of confidence, providing an estimate for a parameter. Let's calculate our data's confidence interval for the mean for 95% confidence level in R.

```r
# calculating point estimations and confidence intervals, for a 95% confidence interval
n <- length(leab) # number of elements
xbar <- mean(leab) # mean
s <- sd(leab) # sample standard deviation

margin <- qt(0.975, df= n-1)*s / sqrt(n) # 3.082897

low <- xbar-margin # lower = 67.57996

high <- xbar+margin # upper = 73.74575
```

Figure 2.2.1: R implementation to calculate confidence interval.

The confidence level seems to be 0.975, but actually a = 0.05 and confidence level is 0.95 since it covers both sides. Margin is calculated as 3.08, lower boundary is 67.58 and upper boundary is 73.75.

3. HYPOTHESES AND TESTS

We know that the mean of sample with 35 elements is 70.66. So, our claim is mean of population is lower than 75. Let's observe if is there enough evidence to support our claim at a= 0.05. We should use t-test, because we don't know the standard deviation of population and also are working on sample.

$$H_0: \mu = 75\,, H_1: \mu < 75 \text{ (claim)}$$

```
> t.test(leab, mu=75, alternative = "less")

        One Sample t-test

data:  leab
t = -2.859, df = 34, p-value = 0.003604
alternative hypothesis: true mean is less than 75
95 percent confidence interval:
     -Inf 73.22797
sample estimates:
mean of x
 70.66286
```

Figure 3.0.1: t-test applied to hypothesis in R.

Since p-value<0.01, we reject the null hypothesis. Result is strongly supporting our claim.

Our second hypothesis is the mean of gross national income per capita is $20000 in the world. Our sample mean is 16196.97. Let's implement the function for 95% confidence interval.

$$H_0: \mu = 20000(claim), H_1: \mu \neq 20000$$

```
> t.test(gni,mu=20000)

        One Sample t-test

data:  gni
t = -1.3167, df = 34, p-value = 0.1968
alternative hypothesis: true mean is not equal to 20000
95 percent confidence interval:
 10327.07 22066.87
sample estimates:
mean of x
 16196.97
```

Figure 3.0.2: t-test applied to hypothesis in R.

Since p-value is 0.1968, we reject the null hypothesis. There is enough evidence to reject the claim.

4. CORRELATION AND LINEAR REGRESSION MODEL

Our problem is " Is there a relationship between life expectancy at birth and GNI per capita?". Are they related? If they are, what is the strength of the relationship? First let's draw a scatter plot. The graph is plotted in R and the codes are given in the figure below.

```
3
4  x <- mySample$GNI
5  y <- mySample$LEAB
6  # Plot with main and axis titles
7  # Change point shape (pch = 19) and remove frame.
8  #  And Add regression line
9  plot(x, y, main = "Scatter Plot",
10      xlab = "Gross National Income", ylab = "Life expected at birth",
11      pch=19, frame = FALSE)
12 abline(lm(y ~ x, data = mtcars), col = "blue")
```

Figure 4.0.1: The codes to draw scatter plot in R.

Life expected at birth is set to y-axis and GNI per capita is set to x-axis. Also, a regression line is added. Now, let's see the plot.

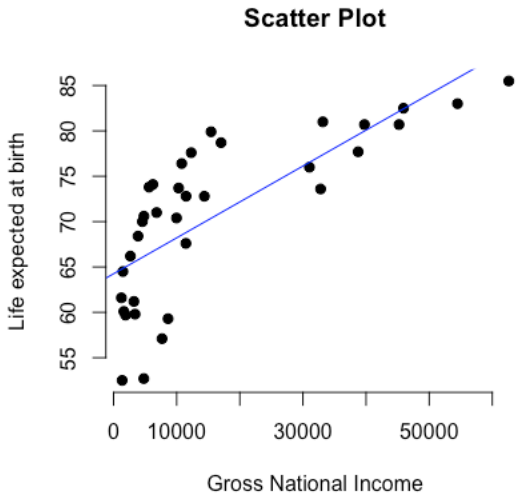


Figure 4.0.2: The scatter plot created in R.

There is a positive linear relationship between them. To strengthen the evidence, let's calculate correlation coefficient of our sample. We do this so that it measures the strength and direction of the linear relationship. The formula for the correlation coefficient is

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

when r is positive, the line slopes upward and to the right. When is negative, it slopes upward and to the right. Also, if there is a strong positive linear relationship between variables, the value will be close to +1. If there is strong negative linear relationship, then value of r will be close to -1.

```
> cor.test(mySample$LEAB, mySample$GNI)

        Pearson's product-moment correlation

data:  mySample$LEAB and mySample$GNI
t = 6.5854, df = 33, p-value = 1.741e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5612864 0.8686765
sample estimates:
      cor
0.7535785
```

Figure 4.0.3: Calculating correlation coefficient in R.

The correlation coefficient is approximately 0.7536 in our experiment. Now, we're going to do hypothesis testing,

$$H_0: p = 0, H_1: p \neq 0$$

null hypothesis means that there is no correlation between variables in the population and alternative hypothesis means that there is significant correlation between the variables in the population.

First we have to apply t-Test for the Correlation Coefficient.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

```
r <- 0.7536
n <- 33
t = r*sqrt((n-2)/(1-r**2))
cv <- qt(0.95,33)
```

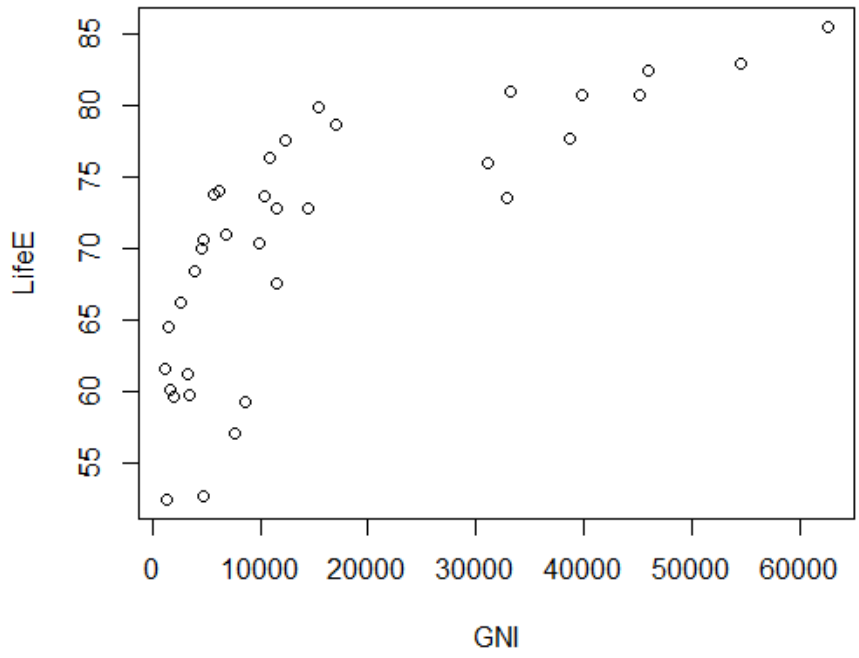
Figure 4.0.4: Implementing in R

t value is 6.383 and critical value is 1.692. This means we reject the null hypothesis and there is correlation between life expectancy at birth and GNI per capita.

The result is there is a positive linear relationship between life expectancy at birth and GNI per capita and it is supported by evidences.

5. ANOVA

```
3  LifeE <- c(proje$LifeE)
4  GNI <- c(proje$GNI)
5  df <- data.frame(LifeE, GNI)
6  plot(LifeE ~ GNI, data = df)
7  proje.aov <- aov(LifeE ~ GNI, data = df)
8  summary(proje.aov)
```



```
> summary(proje.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
GNI          1   1555  1555.1   43.37 1.74e-07 ***
Residuals   33   1183    35.9
---
```

In conclusion, according to the results of this ANOVA analysis, it can be said that there is a statistically significant difference between the groups in the "GNI" factor. The P value is extremely small (1.74e-07), indicating that the "GNI" factor is too large to explain the difference between groups by randomness.

5.1 Two Way ANOVA

```
result <- aov(Life_Expectancy ~ GNI, data = data)
data <- data.frame(GNI = gni_data, Life_Expectancy = life_expectancy_data)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
GNI          1   1555  1555.1   43.37 1.74e-07 ***
Residuals   33   1183    35.9
```

6. Nonparametic Tests

6.1 Single-Sample Sign Test

Null hypothesis=72.8

```
> x=c(82,5, 73.7 ,73.8, 81,80.7 ,70.4 ,78.7 ,70 ,70.6 ,77.7 ,76.4 ,77.6,74.1 ,64.5 ,73.
6, 52.5 ,59.7 ,71 ,52.7 ,72.8 ,59.8 ,85.5 ,61.2 ,72.8 ,80.7 ,66.2 ,60.1 ,68.4 ,83 ,57.1
,76 ,59.3 ,61.6 ,79.9 ,67.6 )
> SIGN.test(x,md=72.8)

        One-sample Sign-Test

data:  x
s = 16, p-value = 0.8642
alternative hypothesis: true median is not equal to 72.8
95 percent confidence interval:
 67.01410 74.89515
sample estimates:
median of x
       71.9
```

Fail to reject null hypothesis

6.2 Paired-Sample Sign Test

For α=0.05 and y is a random sample from our population

```
> y=c(73,83.2,82.7,67.1,84.5,81.4,69.4,82,59.3,81.7,82,82.8,81.9,67.1,82.7,72.6,82.6,5
8.6,84.4 ,60.3,77.2,70.7,70,80.7,81.6,78.7,83,82.5,53.1,82.9,77.1,77.7,80.1,70,61.6)
```

```
> SIGN.test(x,y)

        Dependent-samples Sign-Test

data:  x and y
S = 10, p-value = 0.01667
alternative hypothesis: true median difference is not equal to 0
95 percent confidence interval:
 -10.720811  -1.379189
sample estimates:
median of x-y
         -7.8
```

Fail to reject null hypothesis(p val>0.05)

6.3 Wilcoxon rank-sum Test

For α=0.025 and z is a random sample from our population

```
> z=c(84,83.2,82.7,85.5,84.5,81.4,83,82,59.3,81.7,82,82.8,81.9,67.1,82.7,83.3,82.6,58.
6,84.4 ,60.3,77.2,70.7,83.8,80.7,81.6,78.7,83,82.5,81.2,82.9,77.1,77.7,80.1,70,61.6)
> wilcox.test(x,z,alternative = "greater")

        Wilcoxon rank sum test with continuity correction

data:  x and z
W = 272.5, p-value = 1
alternative hypothesis: true location shift is greater than 0
```

P value>0.025, fail to reject null hypothesis

6.4 Spearman Rank Correlation Coefficient

For α=0.05 and y is a random sample from our population.

```
> cor.test(x,y,method = "spearman")

        Spearman's rank correlation rho

data:  x and y
S = 8361.5, p-value = 0.3258
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1710803
```

Fail to reject null hypothesis

6.5 The Runs Test

```
> runs.test(x)

        Runs Test

data:  x
statistic = 0.1824, runs = 18, n1 = 16, n2 = 17, n = 33, p-value =
0.8553
alternative hypothesis: nonrandomness
```

7. RESOURCES

https://www.kaggle.com/datasets/rajkumarpandey02/human-development-index-and-components?resource=download - The data we applied to our experiment.

https://www.tutorialspoint.com/r/r_csv_files.htm - We used their codes in data collection part.

https://www.datamentor.io/r-programming/histogram - We inspired from their codes about drawing histogram.

https://www.r-bloggers.com - General Information about Confidence Interval and Correlation Coefficient

https://www.geeksforgeeks.org - General Information about R functions.